DISCRETE AND CONTINUOUS MODELS AND APPLIED COMPUTATIONAL SCIENCE

Volume 29 Number 3 (2021)

Founded in 1993
Founder: PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA

DOI: 10.22363/2658-4670-2021-29-3

Edition registered by the Federal Service for Supervision of Communications,
Information Technology and Mass Media
Registration Certificate: ПИ № ФС 77-76317, 19.07.2019

ISSN 2658-7149 (online); 2658-4670 (print)
4 issues per year.
Language: English.

Publisher: Peoples' Friendship University of Russia (RUDN University). Indexed in Ulrich's Periodicals Directory (http://www.ulrichsweb.com), in https://elibrary.ru, EBSCOhost (https://www.ebsco.com), CyberLeninka (https://cyberleninka.ru).

Aim and Scope

Discrete and Continuous Models and Applied Computational Science arose in 2019 as a continuation of RUDN Journal of Mathematics, Information Sciences and Physics. RUDN Journal of Mathematics, Information Sciences and Physics arose in 2006 as a merger and continuation of the series "Physics", "Mathematics", "Applied Mathematics and Computer Science", "Applied Mathematics and Computer Mathematics".

Discussed issues affecting modern problems of physics, mathematics, queuing theory, the Teletraffic theory, computer science, software and databases development.

It's an international journal regarding both the editorial board and contributing authors as well as research and topics of publications. Its authors are leading researchers possessing PhD and PhDr degrees, and PhD and MA students from Russia and abroad. Articles are indexed in the Russian and foreign databases. Each paper is reviewed by at least two reviewers, the composition of which includes PhDs, are well known in their circles. Author's part of the magazine includes both young scientists, graduate students and talented students, who publish their works, and famous giants of world science.

The Journal is published in accordance with the policies of COPE (Committee on Publication Ethics). The editors are open to thematic issue initiatives with guest editors. Further information regarding notes for contributors, subscription, and back volumes is available at http://journals.rudn.ru/miph.

E-mail: miphj@rudn.ru, dcm@sci.pfu.edu.ru.

EDITORIAL BOARD

Editor-in-Chief

Vice Editor-in-Chief

Members of the editorial board

Computer Design: *A. V. Korolkova, D. S. Kulyabov*

Address of editorial board:
Ordzhonikidze St., 3, Moscow, Russia, 115419
Tel. +7 (495) 955-07-16, e-mail: publishing@rudn.ru
Editorial office:
Tel. +7 (495) 952-02-50, miphj@rudn.ru, dcm@sci.pfu.edu.ru
site: http://journals.rudn.ru/miph

Paper size 70×1 00/16. Offset paper. Offset printing. Typeface "Computer Modern".
Conventional printed sheet 6.77. Printing run 500 copies. Open price. The order 616.
PEOPLES' FRIENDSHIP UNIVERSITY OF RUSSIA
6 Miklukho-Maklaya St., 117198 Moscow, Russia
Printed at RUDN Publishing House:
3 Ordzhonikidze St., 115419 Moscow, Russia,
Ph. +7 (495) 952-04-41; e-mail: publishing@rudn.ru

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)
2021, 29 (3) 204
http://journals.rudn.ru/miph

Contents

Viktor V. Andreev, Denis V. Chuprov, Modeling and design of an reconfigurable isolated remote for plasma experiments with hard-real-time synchronization 205

Anatoly Y. Botvinko, Konstantin E. Samouylov, Evaluation of the firewall influence on the session initiation by the SIP multimedia protocol 221

Anatoly Y. Botvinko, Konstantin E. Samouylov, Evaluation of firewall performance when ranging a filtration rule set 230

Tatiana V. Rykova, Towards the analysis of the performance measures of heterogeneous networks by means of two-phase queuing systems . . 242

Aleksandr A. Belov, Valentin S. Khokhlachev, Asymptotically accurate error estimates of exponential convergence for the trapezoidal rule 251

Aleksandr A. Belov, Maxim A. Tintul, Shifted Sobol points and multigrid Monte Carlo simulation 260

Ali Baddour, Mikhail D. Malykh, Richardson–Kalitkin method in abstract description 271

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2021, 29 (3) 205–220
http://journals.rudn.ru/miph

Research article

UDC 53.087.4:533.9.07
PACS 52.59 Ye, 07.05 Dz
DOI: 10.22363/2658-4670-2021-29-3-205-220

Modeling and design of an re-configurable isolated remote for plasma experiments with hard-real-time synchronization

Viktor V. Andreev, Denis V. Chuprov

Peoples' Friendship University of Russia (RUDN University)
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation

(received: July 24, 2021; accepted: September 9, 2021)

The purpose of this paper is to present the design and implementation of a reconfigurable remote control for performing plasma experiments with Hard-Real-Time (HRT) synchronization under jitter less than 1 microsecond. An additional requirement for a multichannel synchronization system is the use of high-speed optical converters to provide galvanic isolation between powerful modules of the setup and remote control in order to exclude any possibility of disruption of the physical experiment control system.

Modeling and development of the software part of the maser remote control panel was performed in the LabVIEW application development environment with Real Time and FPGA modules.

The hardware part of the control panel is implemented on a real-time controller working in conjunction with the Xilinx FPGA module. To ensure the optical isolation of synchronization signals, boards of electron-optical converters based on LED lasers with fiber-optic terminals were developed and manufactured.

The control program is implemented in a two-module architecture with a HOST application and an FPGA application that exchange data over a 1000BASE-T Ethernet network.

Key words and phrases: remote control, synchronization, hard real-time system, FPGA, reconfigurable input-output (RIO)

1. Introduction

Control of a complex multi-parameter physics experiment places high demands on the synchronization of the operation of various systems, nodes and modules of the experimental setup [1]–[3]. In such case, controllability is defined as the ability of the remote control system to achieve a define state when several processes work together or in a certain experiment scenario in a time sequence [4]. Real-time systems must accomplish executive and application tasks within specified timing constraints [5], [6]. The time resolution —

the minimum distinguishable step along the time axis-is in this case the most important characteristic of the synchronization system. So simultaneity of processes or phenomena is reduced to the fact that these processes on the time axis are separated from each other by intervals that do not exceed the time resolution of the system.

The use in physics experiment high-voltage AC and DC or UHF power systems generates a high level of electromagnetic noise and in this case the concept of Power-over-Fiber technology is applied to ensure controllability [7], [8]. Thus, the safety regulations for equipment operating at high power levels, along with the serious requirements for the reliability of the electronics in harsh environments, require galvanic isolation [9], [10].

Our work on synchronization in HRT systems has been performed as part of an ongoing autoresonance plasma experiments in Plasma physics laboratory of RUDN [11], [12]. To organize the operation of the synchronization system, a single time reference point is selected, and the characteristic stages of each of the processes are separated from this beginning by an adjustable amount of delay. The device that provides a countdown of the required delay was called a delay generator. The developed concept can easily be scaled practically on any type of complicated plasma experiments.

2. Experimental setup

As an example of a complex physical installation that requires control with real time synchronization, consider the problem of controlling a plasma maser. A plasma maser (plasma relativistic microwave generator) is a source of powerful microwave radiation in which the Cherenkov effect of the interaction of a high-current relativistic electron beam (REB) with a slow wave of a plasma waveguide is realized [13]–[15]. The configurations of plasma masers differ in the relative position of the REB and plasma and are described in detail in [16], [17].

In the maser scheme implemented in this project, the REB propagates in a strong longitudinal magnetic field in a metal cylindrical waveguide, in which a plasma with controlled parameters is pre-created.

A simplified block diagram of the created plasma maser is shown in the figure 1 and includes an main experimental setup and a system for ensuring its operation.

The main experimental setup includes 4 modules:

- A generator of periodic high-voltage pulses of nanosecond duration as a source of a high-current REB in a direct-acting accelerator based on an explosive-emission cathode.
- A plasma source with controlled parameters.
- The space of formation of REB and plasma in a strong magnetic field.
- Remote control of all nodes and diagnostic systems.

The support system consists of:

- A vacuum pumping system.
- Gas inlet system for creating plasma.
- A system for converting a plasma wave into an electromagnetic wave, releasing radiation into the atmosphere, focusing it, and transporting it.
- Radiation detection and diagnostics system.

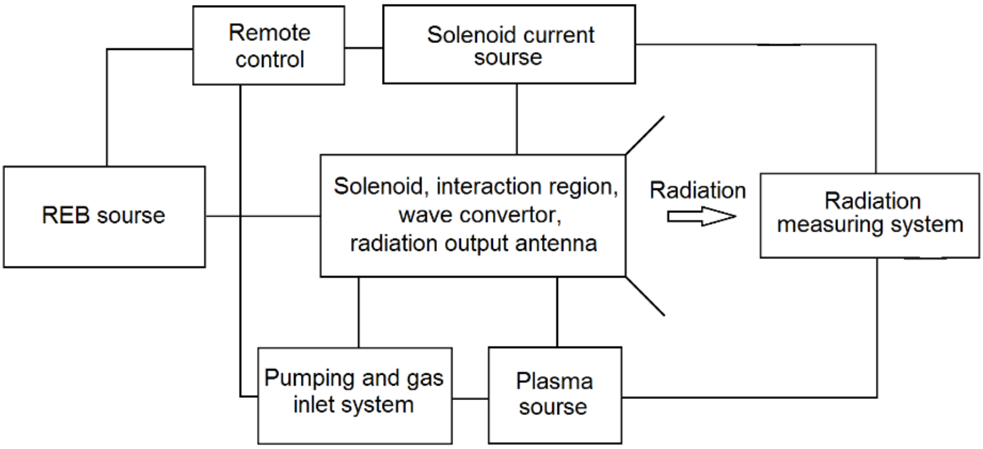

Remote control
Solenoid current sourse
REB sourse
Solenoid, interaction region, wave convertor, radiation output antenna
Radiation
Radiation measuring system
Pumping and gas inlet system
Plasma sourse

Figure 1. Block diagram of a plasma maser

The operation of all the main operating systems of the plasma maser is provided by a synchronization system with a strict reference to real time. This system is integrated into the control panel, where the operator sets the required operation scenario, provides a physical start of the maser and receives a report in the form of tabular and graphical data on the state of the maser subsystems at synchronous times of the working cycle. Schematically, the maser operation scenario is shown in the figure 2.

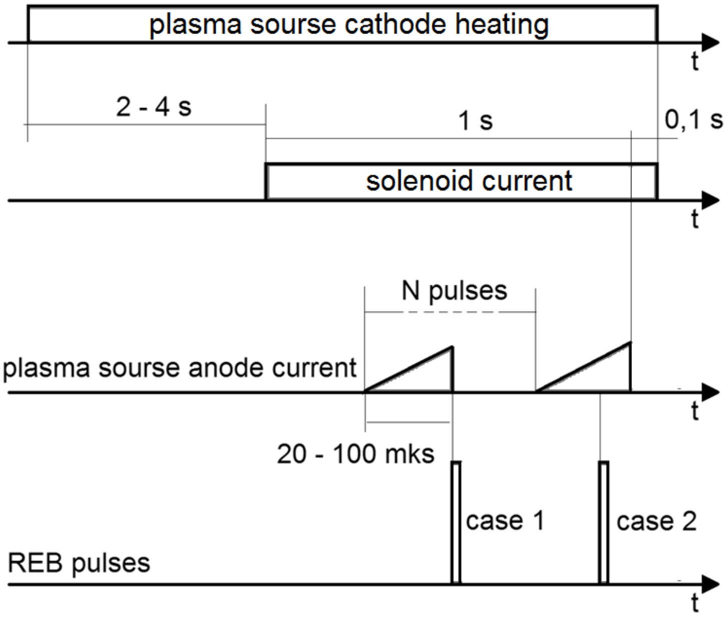


Figure 2. Plasma maser working scenario

If the warm solenoid is ready, after 30 seconds after all the limit switches (locks) are triggered, a ready signal is generated — the “Start” button in the remote control interface is activated. Pressing the “Start” button generates a “glow” pulse, which is applied to the power supply of the plasma source.

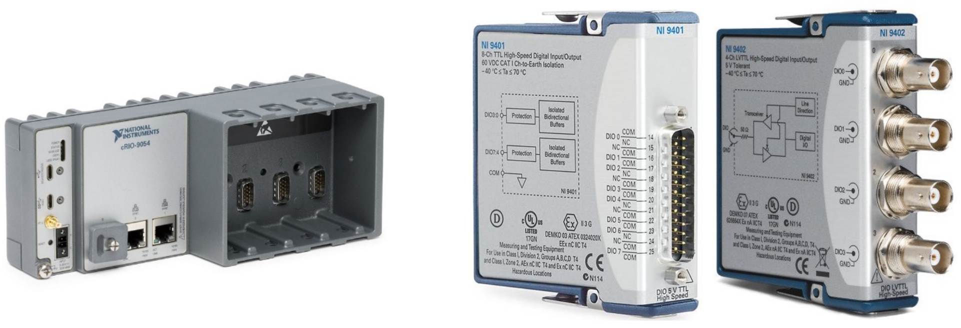
After the set time “Glow duration” (set by the operator in the range of 0–5 s), the signal “solenoid current” is generated. After the battery charge current of the solenoid reaches the set values, the remote control generates a burst of pulses with the set frequency during 1 s.

The developed remote control allows the operator to implement two options for starting the maser. In case 1 (see the figure 2) a beam of relativistic electrons interacts with a “quiet” plasma in the absence of current in the cathode circuit of the plasma generator. The plasma decay time is of the order of tens to hundreds of microseconds, so synchronization with a jitter of the order of 1 microsecond is enough. In the development of the launch scenario according to the case 2 option, the REB interacts with the plasma under conditions of increasing concentration. One of the tasks of the experiment is to compare the two described scenarios and identify the preferred conditions for the occurrence and development of a plasma-beam discharge in terms of obtaining a powerful broadband EMR pulse.

After a time of about 1 ms after the last pulse of the bundle, the remote control generates the endings (trailing edges) of the current pulses of both warm solenoids and the glow pulse of the thermocathode.

The developed remote provides control of the signals of the security system, as well as the generation of appropriate enabling, warning, or prohibiting signals and commands.

The NI cRIO-9053 chassis (see the figure 3(a)) with a real-time controller and an integrated field-programmable gate array (FPGA) chip was chosen as the hardware platform of the synchronization system. The FPGA architecture is a set of programatically configurable logic blocks, the connections between them, and the I/O blocks. This structure is best suited to the tasks of parallel multi-channel data processing to multi-channel signal generation.



(a)

(b)

Figure 3. NI cRIO-9053 chassis (a); I/O modules NI 9401 and NI 9402 (b)

The reconfigurable I/O modules NI 9401 and NI 9402 were used for matching with the generator loads — input channels of synchronized devices (see the figure 3(b)). They have a similar circuit architecture but differ in the maximum switching speed and the form factor of the output connectors. This solution is chosen to conditionally divide the input and output signals into two groups: “fast” signals with strict jitter requirements (starting the plasma generation and starting the REB source) and “slow” signals with a relatively less rigid reference to real time (starting the heating of the thermal cathode, starting the solenoid current generator and emergency lock signals).

Main characteristics of the controller and modules:

— 4-slot CompactRIO controller, 1.30 GHz Dual-core CPU, 4 GB DRAM, 4 GB Storage, −20°C to 55°C, Artix-7 50T FPGA. The cRIO-9053 controller is a secure, high-performance, customizable embedded controller that contains a dual-core Intel Atom processor, an Artix-7 FPGA, and four slots for C-series modules. It runs on the NI Linux Real-Time operating system with I/O access via the NI-DAQmx drivers or via the LabVIEW FPGA module.

— The NI 9401 module provides a reconfigurable I/O interface for digital lines in 4-bit increments and operates in three configurations: 8 digital inputs, 8 digital outputs, or 4 digital inputs and outputs each. The pulse font is not worse than 100 ns. The signal level is 5 V TTL. Output connector 25 PIN D-SUB.

— The NI 9402 module provides a reconfigurable I/O interface for 4 digital lines, operating in two configurations: 4 digital inputs, or 4 digital outputs. The pulse font is not worse than 50 ns. The signal strength is 3.3 V LVTTL. 4X BNC (50 Ohm) output connector.

A digital 4-channel RIGOL MSO1074 oscilloscope with a bandwidth of 70 MHz and a minimum detectable pulse duration of 10 ns was used to monitor the system and perform test measurements.

3. Results and conclusion

A program of control of the prototype generator was created in the LabVIEW graphical programming environment [18]–[20]. LabVIEW is one of the most popular development environments for modeling, simulation and equipment control applications.

The developed program consists of two parts: the top-level host program in the figure 4 is responsible for the user interface and transmits the synchronization system settings and setpoints to the FPGA program once per second. The current values of the settings are transmitted over the communication channel between the HOST and the FPGA applications.

In the FPGA program in the figure 5 after checking the health of all connected devices and assigning program identifiers to the physical synchronization channels, an infinite loop begins, working out the commands of the HOST application. After receiving the “START” command, the synchronization signals are switched on and off sequentially with the specified durations and delays for each of the active channels. After the plasma maser cycle is completed, the FPGA application signals this to the HOST application and goes back to standby mode.

The waveforms with the results of testing the program of the remote control are shown in the figure 6. Beam 1 — the starting pulse, beam 2 — the output pulse of one of the synchronization channels (starting the plasma generation). Tests have shown that the programmable delay of the generated signal corresponds to the setpoint with high accuracy.

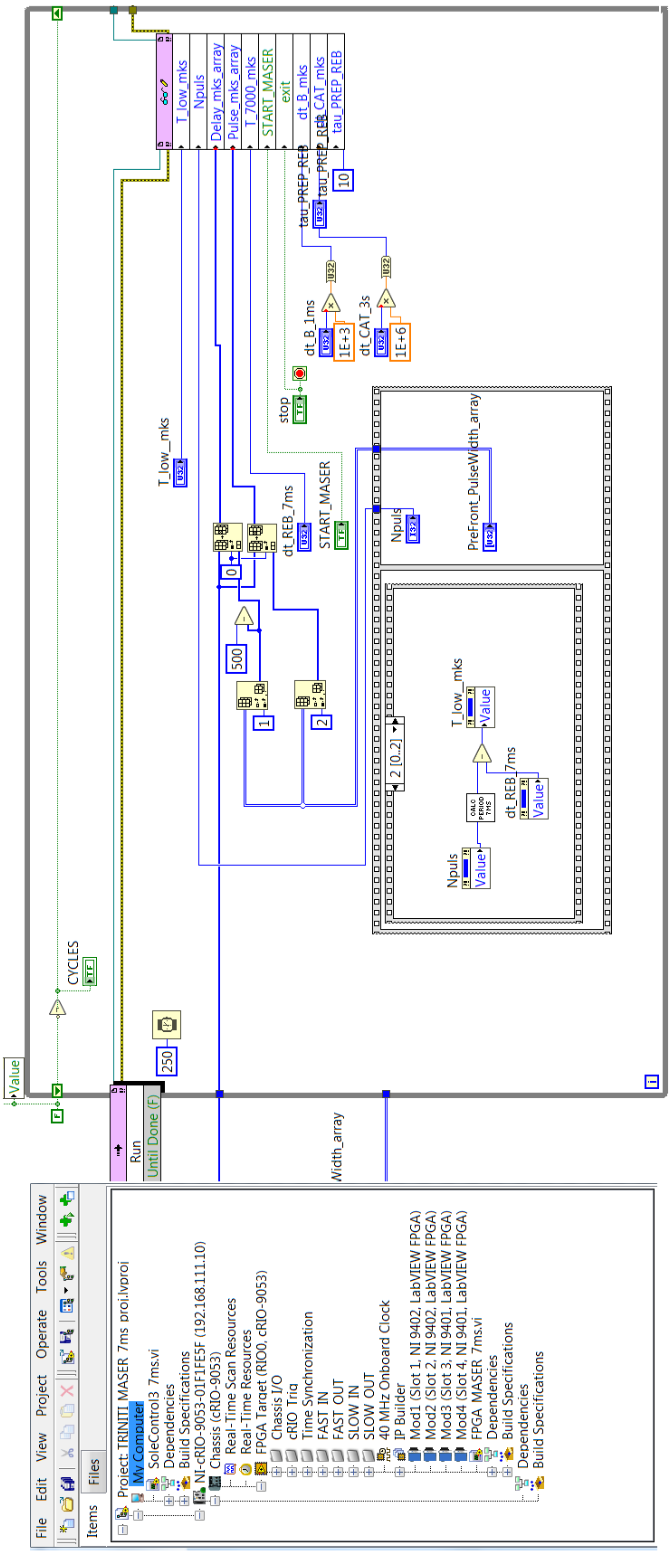


Figure 4. The project tree and a part of the HOST application

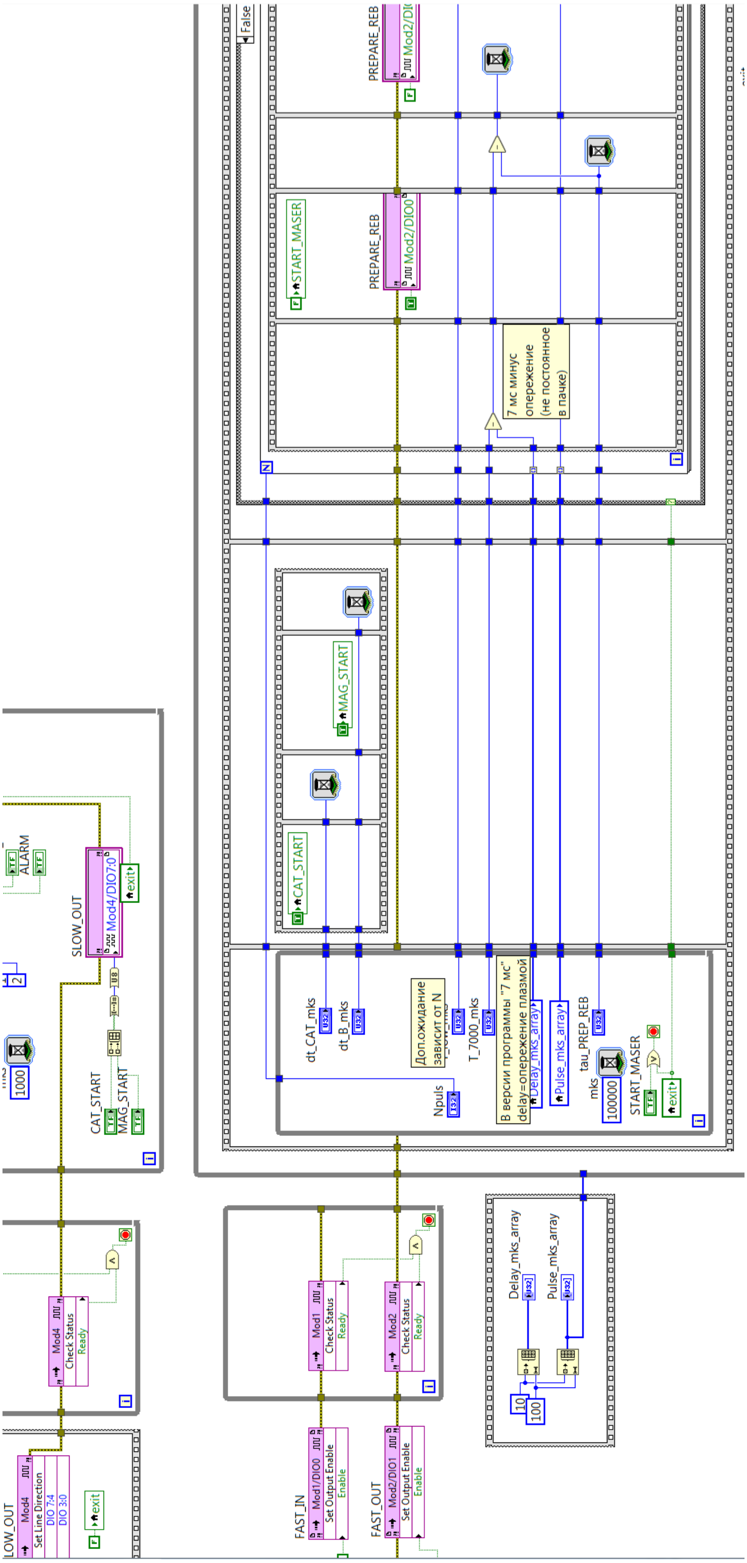
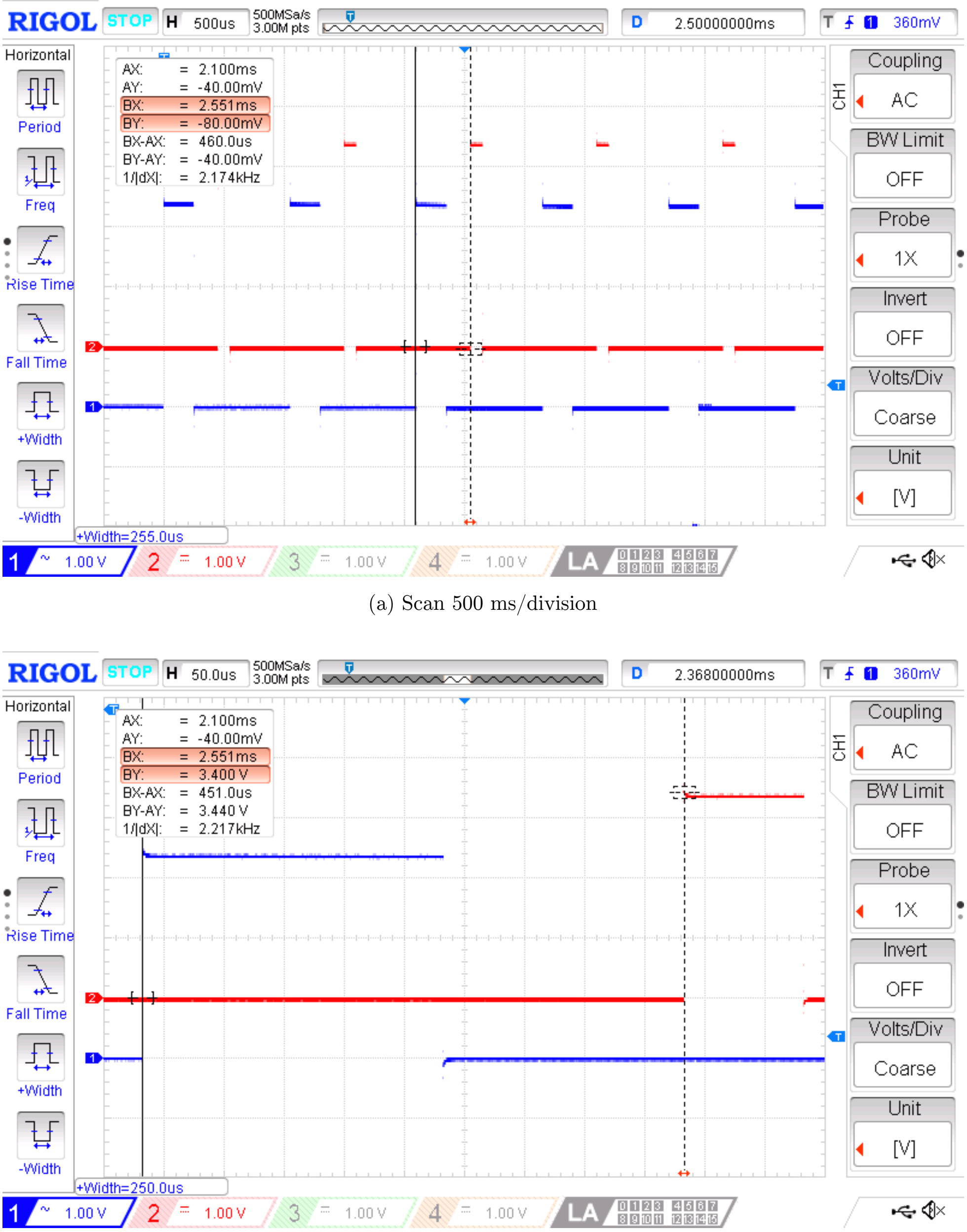


Figure 5. Part of the FPGA application

The jitter observed at the signal edges is less than 100 ns (it is not noticeable on the waveforms). The step of changing the programmable delay in the synchronization system channel is 1 microsecond. The delay can vary from 1 microsecond to 2^{32} microseconds (more than 1 hour).



H 500us 500MSa/s 3.00M pts
D 2.50000000ms
T 360mV

AX: = 2.100ms
AY: = -40.00mV
BX: = 2.551ms
BY: = -80.00mV
BX-AX: = 460.0us
BY-AY: = -40.00mV
1/|dX|: = 2.174kHz

Coupling AC
BW Limit OFF
Probe 1X
Invert OFF
Volts/Div Coarse
Unit [V]

+Width=255.0us

1 1.00 V 2 1.00 V 3 1.00 V 4 1.00 V

(a) Scan 500 ms/division

H 50.0us 500MSa/s 3.00M pts
D 2.36800000ms
T 360mV

AX: = 2.100ms
AY: = -40.00mV
BX: = 2.551ms
BY: = 3.400 V
BX-AX: = 451.0us
BY-AY: = 3.440 V
1/|dX|: = 2.217kHz

Coupling AC
BW Limit OFF
Probe 1X
Invert OFF
Volts/Div Coarse
Unit [V]

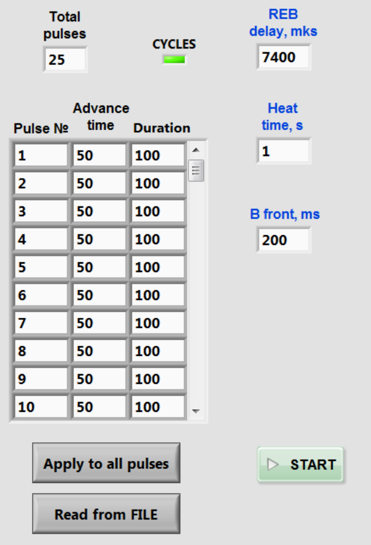
+Width=250.0us

1 1.00 V 2 1.00 V 3 1.00 V 4 1.00 V

(b) Scan 50 ms/division

Figure 6. Waveforms of the operation of the plasma generation start channel in single-channel mode with a delay time setting of 450 microseconds

The next stage of our tests was to check the software control of the delays in the generation of the REB relative to the front of the plasma pulse under the conditions of the generation of a pulse train. To configure program delays, the control program interface provides a function block, shown in the figure 7.



Total pulses: 25

REB delay, mks: 7400

Heat time, s: 1

B front, ms: 200

Pulse №	Advance time	Duration
1	50	100
2	50	100
3	50	100
4	50	100
5	50	100
6	50	100
7	50	100
8	50	100
9	50	100
10	50	100

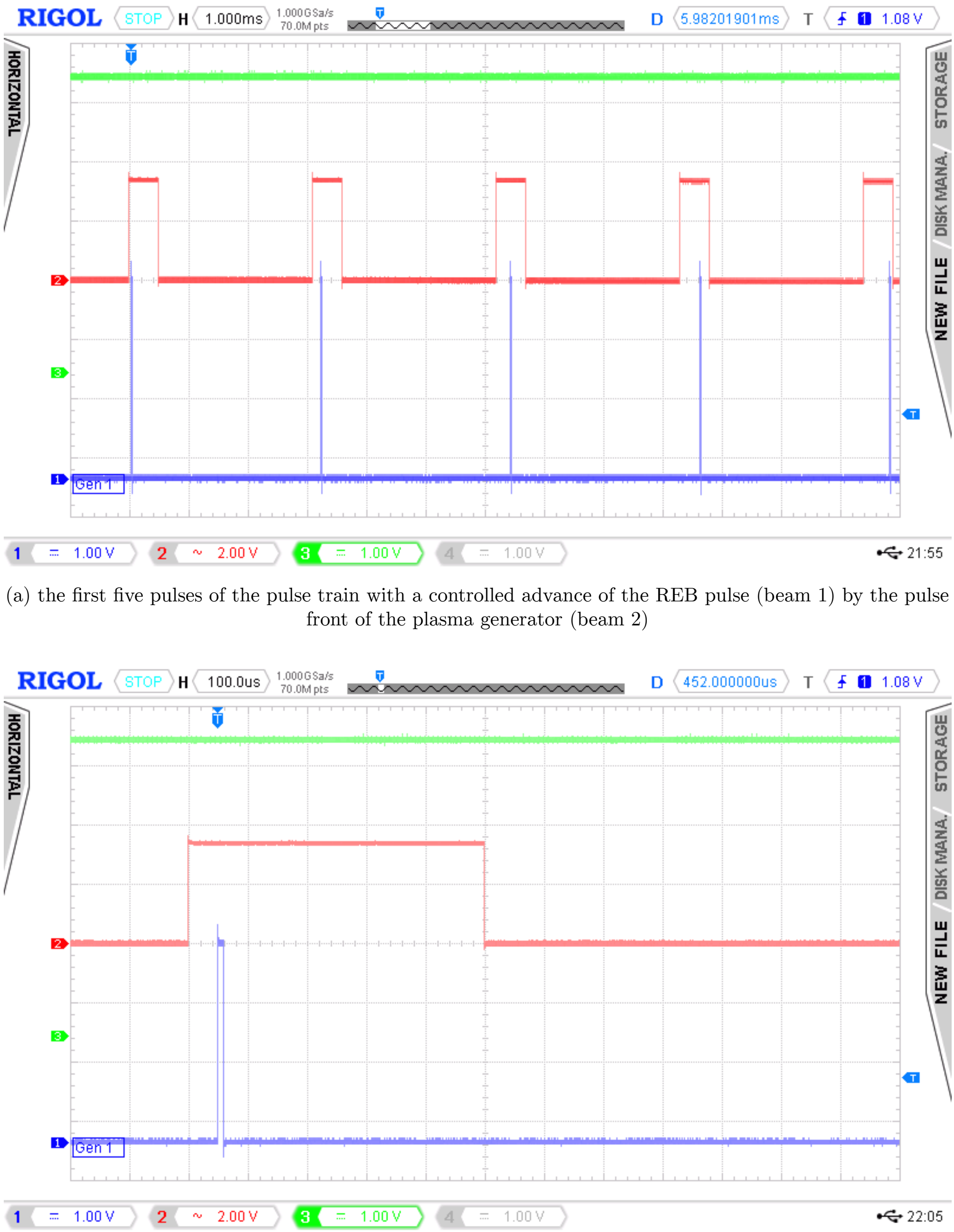
Figure 7. The control program interface block, with channel delay settings

After specifying the number of pulses in the packet, the values of the plasma pulse durations and lead times (by how many microseconds the front of the plasma pulse is ahead of the RAP pulse) should be entered in the program delay table for each pulse of the packet. The settings made in the first row can be applied to the entire table. It is possible to read the corresponding table from a text file in a similar format to the table. The separator is a Tab character, the line separators are the End Of Line character.

A series of waveforms showing the possibilities of software tuning of the advance is shown in figures 8, 9, 10. The beam 2 is a synchropulse for plasma generation, the beam 1 is a program repeater of the REB pulse with a duration of 10 microseconds (for easy display on oscillograms).

It can be seen that the leading edge of the plasma pulse of the REB pulse strictly follows the setpoints of the control program interface (see the figure 7).

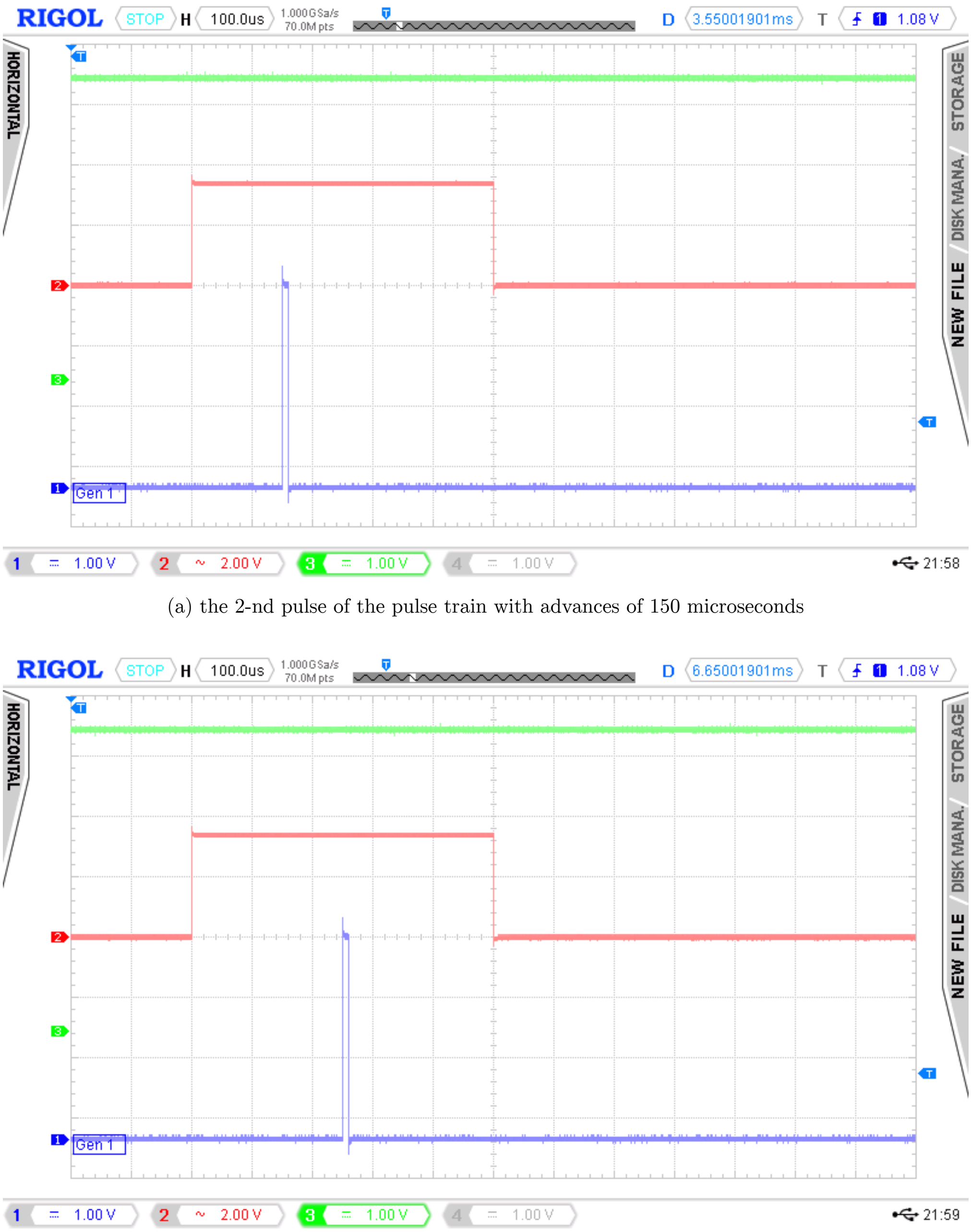
The created software and hardware complex allows you to scale the number of synchronization channels and, with the available equipment, get a total of 8 delay channels with edges no worse than 50 ns and 16 channels with edges no worse than 100 ns. The control panel based on the implemented synchronization system has shown the ability for long-term stable uninterrupted operation and is currently successfully used in experiments to select the optimal operating modes of a plasma maser. Additionally, if it is necessary to switch to a submicrosecond delay control step, the system can be reconfigured to work with synchronization directly from the FPGA clock generator. The existing FPGA platform with a clock frequency of 40 MHz allows you to reduce the minimum step of regulating pulse durations and delays to 25 ns. It should, however, be taken into account that the control range in this case will decrease by a multiple to a value of about 15 minutes.



(a) the first five pulses of the pulse train with a controlled advance of the REB pulse (beam 1) by the pulse front of the plasma generator (beam 2)

(b) the 1-st pulse of the pulse train with advances of 50 microseconds

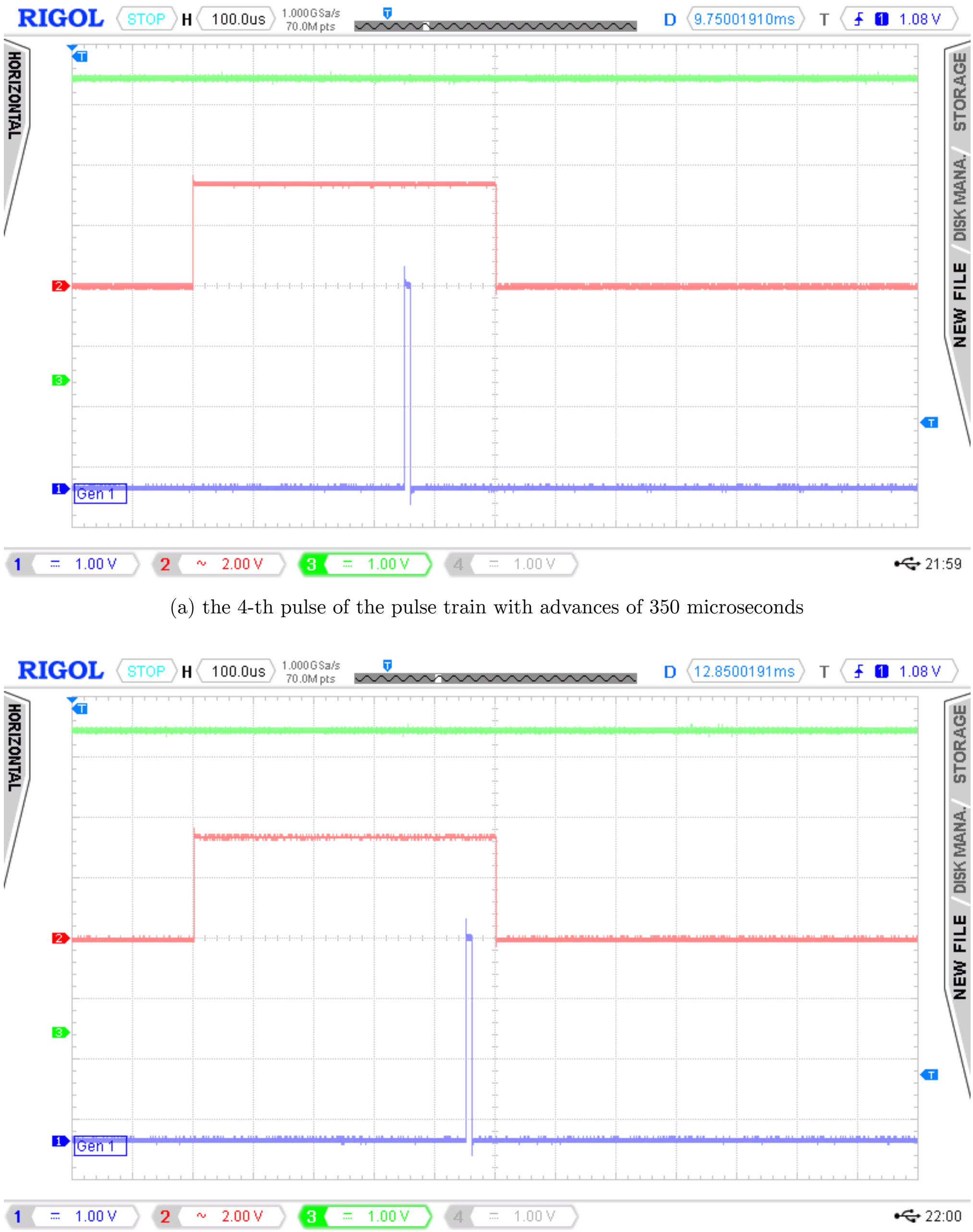
Figure 8. The possibilities of tuning the advance between channels in the pulse train generation mode



(a) the 2-nd pulse of the pulse train with advances of 150 microseconds

(b) the 3-rd pulse of the pulse train with advances of 250 microseconds

Figure 9. The possibilities of tuning the advance between channels in the pulse train generation mode



(a) the 4-th pulse of the pulse train with advances of 350 microseconds

(b) the 5-th pulse of the pulse train with advances of 450 microseconds

Figure 10. The possibilities of tuning the advance between channels in the pulse train generation mode

The developed multi-channel synchronization system with fiber-optic galvanic isolation of the controller and secondary circuits of the equipment can be easily adapted to any tasks of controlling a physical experiment. This, in particular, is in demand when conducting experiments with cold plasma in the IFIT RUDN. These experiments require synchronization of the operation of a high-power microwave generator and a pulse current generator, as well as synchronized measurements of radiation parameters from the region of the plasma clot localization in the microwave, optical and X-ray ranges, probe measurements in the low-frequency and microwave ranges, etc. With the help of the created system, it is possible to provide such synchronization with a time resolution of at least 1 microsecond.

Acknowledgement

The reported study was funded by RFBR, project number 18-29-21041 (recipient D.V. Chuprov, Results and conclusion) and partially agreement 226/2480 (recipient V.V. Andreev, experimental setup).

References

[1] K. Patel, N. Umesh, H. C. Joshi, S. Pathak, K. A. Jadeja, K. Patel, and R. L. Tanna, “LabVIEW-FPGA-based real-time data acquisition system for ADITYA-U heterodyne interferometry,” *IEEE Transactions on Plasma Science*, vol. 49, no. 6, pp. 1891–1897, 2021. DOI: 10.1109/TPS.2021.3082159.

[2] M. Kim and M. Kwon, “LabVIEW-EPICS interfaces in KSTAR control system,” in *Proc. 9th Int. Conf. on Accelerator and Large Experimental Physics Control Systems (ICALEPCS’03)*, Paper MP519, Gyeongju, Korea, Oct. 2003, pp. 87–89.

[3] Y. Ege, M. Kabadayi, O. Kalender, M. Coramik, H. Citak, E. Yuruklu, and A. Dalcali, “A new electromagnetic helical coilgun launcher design based on LabVIEW,” *IEEE Transactions on Plasma Science*, vol. 44, no. 7, pp. 1208–1218, 2016. DOI: 10.1109/TPS.2016.2575080.

[4] P. S. Korenev, Y. V. Mitrishkin, and M. I. Patrov, “Reconstruction of equilibrium distribution of Tokamak plasma parameters by external magnetic measurements and construction of linear plasma models [Rekonstruktsiya ravnovesnogo raspredeleniya parametrov plazmy Tokamaka po vneshnim magnitnym izmereniyam i postroyeniye lineynykh plazmennykh modeley],” *Mekhatronika, Avtomatizatsiya, Upravlenie*, vol. 17, 4 2016, in Russian. DOI: 10.17587/mau.17.254-266.

[5] L. Giannone *et al.*, “Real time magnetic field and flux measurements for tokamak control using a multi-core PCI Express system,” in *Proc. 25th SOFT*, Id. Nr. 367, Rostock, Sep. 2008.

[6] J. H. Lee, S. H. Lee, S. H. Son, W. H. Ko, D. C. Seo, I. Yamada, K. H. Her, J. S. Jeon, and M. G. Bog, “Development of prototype polychromator system for KSTAR Thomson scattering diagnostic,” *Journal of Instrumentation*, vol. 10, no. 12, p. C12012, Dec. 2015. DOI: 10.1088/1748-0221/10/12/c12012.

[7] K. Sharifabadi, L. Harnefors, H.-P. Nee, S. Norrga, and R. Teodorescu, *Design, control and application of modular multilevel converters for HVDC transmission systems*. John Wiley & Sons, Ltd., 2016.

[8] G. Anda, D. Dunai, M. Lampert, T. Krizsanóczi, J. Németh, S. Bató, Y. U. Nam, G. H. Hu, and S. Zoletnik, "Development of a high current 60 keV neutral lithium beam injector for beam emission spectroscopy measurements on fusion experiments," *Review of Scientific Instruments*, vol. 89, no. 1, p. 013 503, 2018. DOI: 10.1063/1.5004126.

[9] E. Ragonese, N. Spina, A. Parisi, and G. Palmisano, "An experimental comparison of galvanically isolated DC-DC converters: isolation technology and integration approach," *Electronics*, vol. 10, p. 1186, 2021. DOI: 10.3390/electronics10101186.

[10] C. Budelmann, "Opto-electronic sensor network powered over fiber for harsh industrial applications," *IEEE Transactions on Industrial Electronics*, vol. 65, pp. 1170–1177, 2 2018. DOI: 10.1109/TIE.2017.2733479.

[11] V. V. Andreev *et al.*, "Gyromagnetic autoresonance plasma bunches in a magnetic mirror," *Physics of Plasmas*, vol. 24, no. 9, p. 093 518, 2017. DOI: 10.1063/1.4986009.

[12] V. V. Andreev, A. A. Novitsky, and D. V. Chuprov, "The use of streak photography, X-ray radiography, and radiometric and spectrometric measurements to study plasma bunches generated under gyroresonant interactions," *Physics of Atomic Nuclei*, vol. 82, no. 10, pp. 1404–1413, 2019. DOI: 10.1134/S1063778819100016.

[13] M. V. Kuzelev *et al.*, "Plasma relativistic microwave electronics," *Plasma Physics Reports*, vol. 27, pp. 669–691, 8 2001. DOI: 10.1134/1.1390539.

[14] S. E. Ernyleva, V. O. Litvin, O. T. Loza, and I. L. Bogdankevich, "Promising source of high-power broadband microwave pulses with radiation frequency variable up to two octaves," *Technical Physics*, vol. 59, pp. 1228–1232, 8 2014. DOI: 10.1134/S1063784214080106.

[15] S. E. Ernyleva and O. T. Loza, "Plasma relativistic microwave noise amplifier with inverse configuration [Plazmennyy relyativistskiy SVCH-usilitel' shuma s inversnoy geometriyey]," *Trudy instituta obschey fiziki im. A.M. Prokhorova*, vol. 72, pp. 128–133, 2016, in Russian.

[16] A. B. Buleyko, N. G. Gusein-zade, and O. T. Loza, "Plasma masers: status quo and development prospects," *Physics of Wave Phenomena*, vol. 26, no. 4, pp. 317–322, 2018. DOI: 10.3103/S1541308X18040118.

[17] A. B. Buleyko, A. V. Ponomarev, O. T. Loza, *et al.*, "Experimental plasma maser as a broadband noise amplifier. II: Short pulse," *Physics of Plasmas*, vol. 28, p. 023 304, 2021. DOI: 10.1063/5.0031432.

[18] P. A. Blume, *The LabVIEW Style Book*. NJ: Upper Saddle River, 2007.

[19] P. Ponce-Cruz and F. D. Ramírez-Figueroa, *Intelligent control systems with LabVIEW*. New York: Springer LDH, 2010, p. 216.

[20] S. Hauck and A. DeHon, Eds., *Reconfigurable computing: the theory and practice of FPGA-based computation*. New York: Elsevier Inc., 2008.

For citation:

V. V. Andreev, D. V. Chuprov, Modeling and design of an re-configurable isolated remote for plasma experiments with hard-real-time synchronization, Discrete and Continuous Models and Applied Computational Science 29 (3) (2021) 205–220. DOI: 10.22363/2658-4670-2021-29-3-205-220.

Information about the authors:

Viktor V. Andreev — Candidate of Physical and Mathematical Sciences, Assistant professor of Institute of Physical Research and Technology of Peoples' Friendship University of Russia (RUDN University) (e-mail: andreev-vv@rudn.ru, phone: +7(495) 9550827, ORCID: https://orcid.org/0000-0002-2654-6752, ResearcherID: O-2878-2013, Scopus Author ID: 23014039400)

Denis V. Chuprov — Senior Lecturer of Institute of Physical Research and Technology of Peoples' Friendship University of Russia (RUDN University) (e-mail: chuprov-dv@rudn.ru, phone: +7(495) 9550759, ORCID: https://orcid.org/0000-0002-6768-6196, ResearcherID: O-3193-2013, Scopus Author ID: 6508067157)

УДК 53.087.4:533.9.07
PACS 52.59 Ye, 07.05 Dz
DOI: 10.22363/2658-4670-2021-29-3-205-220

Моделирование и разработка реконфигурируемого пульта управления для плазменных экспериментов с жёсткой синхронизацией в реальном времени

В. В. Андреев, Д. В. Чупров

Российский университет дружбы народов
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия

Цель данной статьи — представить дизайн и реализацию реконфигурируемого пульта дистанционного управления для проведения плазменных экспериментов с синхронизацией в режиме жёсткого реального времени при джиттере менее 1 микросекунды. Дополнительным требованием к системе многоканальной синхронизации является использование высокоскоростных оптических преобразователей для обеспечения гальванической развязки между мощными модулями установки и дистанционного управления, чтобы исключить любую возможность нарушения работы системы управления физическим экспериментом.

Моделирование и разработка программной части пульта дистанционного управления мазером проводились в среде разработки приложений LabVIEW с модулями Real Time и FPGA.

Аппаратная часть панели управления реализована на контроллере реального времени, работающем совместно с модулем Xilinx FPGA. Для обеспечения оптической развязки сигналов синхронизации разработаны и изготовлены платы электронно-оптических преобразователей на основе светодиодных лазеров с оптоволоконными выводами.

Программа управления реализована в двухмодульной архитектуре с приложением HOST и приложением FPGA, которые обмениваются данными по сети 1000BASE-T Ethernet.

Ключевые слова: пульт управления, синхронизация, система жёсткого реального времени, настраиваемый ввод-вывод

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2021, 29 (3) 221–229
http://journals.rudn.ru/miph

Research article

UDC 519.872
DOI: 10.22363/2658-4670-2021-29-3-221-229

Evaluation of the firewall influence on the session initiation by the SIP multimedia protocol

Anatoly Y. Botvinko[1], **Konstantin E. Samouylov**[1, 2]

[1] *Peoples' Friendship University of Russia (RUDN University)*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[2] *Research Center "Computer Science and Control" of the Russian Academy of Sciences*
44-2, Vavilov St., Moscow, 119333, Russian Federation

(received: August 4, 2021; accepted: September 9, 2021)

Firewalls is one of the major components to provide network security. By using firewalls, you can solve such problems as preventing unauthorized access, and deleting, modifying and/or distributing information under protection. The process of information flows filtration by a firewall introduces additional time delays, thus possibly leading to disruption of stable operation of the protected automated system or to inaccessibility of the services provided by the system. Multimedia services are particularly sensitive to service time delays. The main purpose of the work presented in this paper is to evaluate the influence of the firewall on the time delays in data transmission process in the automated system with multimedia data transmission protocols. The evaluation is provided by the queuing theory methods while a session is initiated between two users by the Session Initiation Protocol (SIP) with firewall message filtration. A firewall is a local or functional distributing tool that provides control over the incoming and/or outgoing information in the automated system (AS), and ensures the protection of the AS by filtering the information, i.e., providing analysis of the information by the criteria set and making a decision on its distribution.

Key words and phrases: SIP, firewall, session initiation, queuing system, filtering time, automated system

1. Introduction

Currently, one of the necessary conditions to provide information security of automated systems is to use software and hardware systems that filter incoming and outgoing traffic. Firewalls increase the time delays for information flows while they are checked in the AS. For multimedia protocols, significant time delays can adversely affect QoE and QoS quality indicators [1] and lead to inability of using the multimedia services provided. Therefore, the evaluation of the firewall influence on the time delays in the data transmission process in the AS with multimedia data transmission protocols is an urgent and demanded task.

To evaluate the firewall influence on the data transmission delay in the AS, the most delay-sensitive service has been selected, i.e., the session initiation by the Session Initiation Protocol (SIP). The script is the initiation of a session between two users with proxy servers and firewall packet filtration.

This paper has the following structure. The process of the session initiation by the SIP protocol is described in Section 2. A method for evaluation of temporal characteristics of the session initiation by the SIP protocol is given in Section 3. The results of the evaluation of the firewall influence on the session initiation time and the session request delay are presented in Section 4. The Conclusion contains the main aspects of the study.

2. Session initiation by the SIP protocol in the presence of firewall

The SIP protocol, developed by the MMUSIC group of the IETF committee, provides for three main types of scripts for initiating a connection: by proxy servers, by a redirecting server, and directly between user [2]–[4]. The main difference in these scenarios is the way of searching and inviting the user. These operations are assigned either to the proxy server, or to the redirecting server, or directly to the user if he knows the address of the called subscriber.

To evaluate the firewall influence on the connection initiation by the SIP protocol, without limiting the generality of the approach, the script for initiating a connection between two users with two proxy servers and one firewall located in the middle of the chain has been considered. The network segment with the client's equipment of the 1st user (User 1) is considered to be the AS under protection — this segment is protected by the firewall. The firewall introduces an additional time delay while checking the compliance of the network packet parameters with the filtration rules specified in the AS under protection.

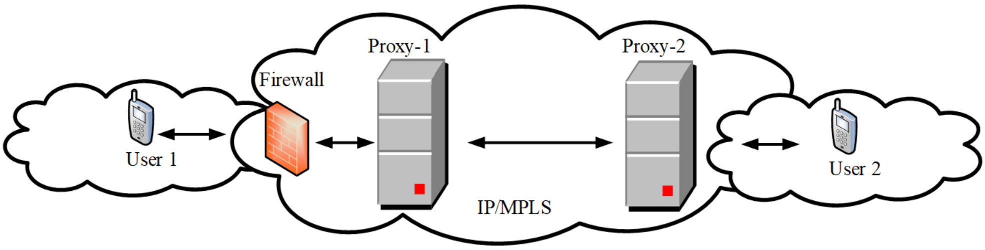


Figure 1. Arrangement of the elements when the SIP session is initiated

The figure 1 shows the elements participating in the connection establishment: user's equipment — User 1, User 2; proxy servers — Proxy-1, Proxy-2; firewall and IP/MPLS main transmission network.

Let's describe the session initiation algorithm, i.e., the sequence of requests and responses of the session initiation process for the script under consideration in accordance with the figure 1.

Session initiation on the equipment of User 1 is Invite message containing the information about the address of the called user — User 2. The message passes through the elements of the firewall and the proxy server, and the element simulating the IP/MPLS network, and the User 2 element. After successful message processing (message retransmission isn't considered), the equipment of User 2 responds with the message 100 Trying. This means that the request is being processed. Then, the equipment of User 2 sends a 180 Ringing message to the User 1. That means that the incoming call signal has been received and the location of the called user has been detected. After processing the Invite request, User 2 generates a 200 Ok response. This response to the Invite request contains the information indicating that the user has agreed to participate in the communication session. The session initiation algorithm is completed by sending the Ack message indicating that the response to the Invite request has been accepted.

Consideration of this session initiation algorithm allows to evaluate the following temporal characteristics of the SIP session initiation service: average session initiation time T_S and average session request delay (SRD) T_{SRD} [5]. T_S is considered from sending the Invite message to the start of the data transmission process. T_{SRD} is considered from the moment the session has been initiated until the first subscriber receives a 180 Ringing response.

The sequence of transmitted signaling messages in the described algorithm of session initiation by SIP protocol is presented in the figure 2 [6].

3. Evaluation of the temporal characteristics of the service of session initiation by SIP protocol in the presence of firewall

To evaluate the firewall influence on the T_S and T_{SRD} times, a mathematical model in the form of an open exponential queuing network (EQN) is proposed [7]. The residence time in the EQN will be equal to the session initiation time [8].

EQN consists of six nodes, each of them modeling a corresponding functional element in the session initiation process. The blocks — User 1, User 2, IP/MPLS — are modeled by the queuing system (QS) $M|M|\infty$, and the rest of the blocks — by the QS $M|M|1|\infty$. Let's introduce the following designation: λ_0 is the intensity of the SIP message flow in the EQN, and μ_i is the service intensity in the i-th node.

So, the condition for the existence of a stationary mode is [9], [10]:

$$\lambda_0 < \min\left(\frac{\mu_2}{5}; \frac{\mu_3}{5}; \frac{\mu_5}{4}\right). \tag{1}$$

Taking into account that the T_S and T_{SRD} times consist of the time of message processing by the functional elements and the waiting time in the queue, and considering the approach given in [5], [8], [9], [11]–[14], we determine the T_{SRD} and T_S times as follows:

$$T_{SRD} = 2\mu_1^{-1} + \frac{2}{\mu_2 - 5\lambda_0} + \frac{2}{\mu_3 - 5\lambda_0} + 2\mu_4^{-1} + \frac{2}{\mu_5 - 4\lambda_0} + \mu_6^{-1}; \tag{2}$$

$$T_S = 2\mu_1^{-1} + \frac{3}{\mu_2 - 5\lambda_0} + \frac{3}{\mu_3 - 5\lambda_0} + 3\mu_4^{-1} + \frac{3}{\mu_5 - 4\lambda_0} + 2\mu_6^{-1}. \tag{3}$$

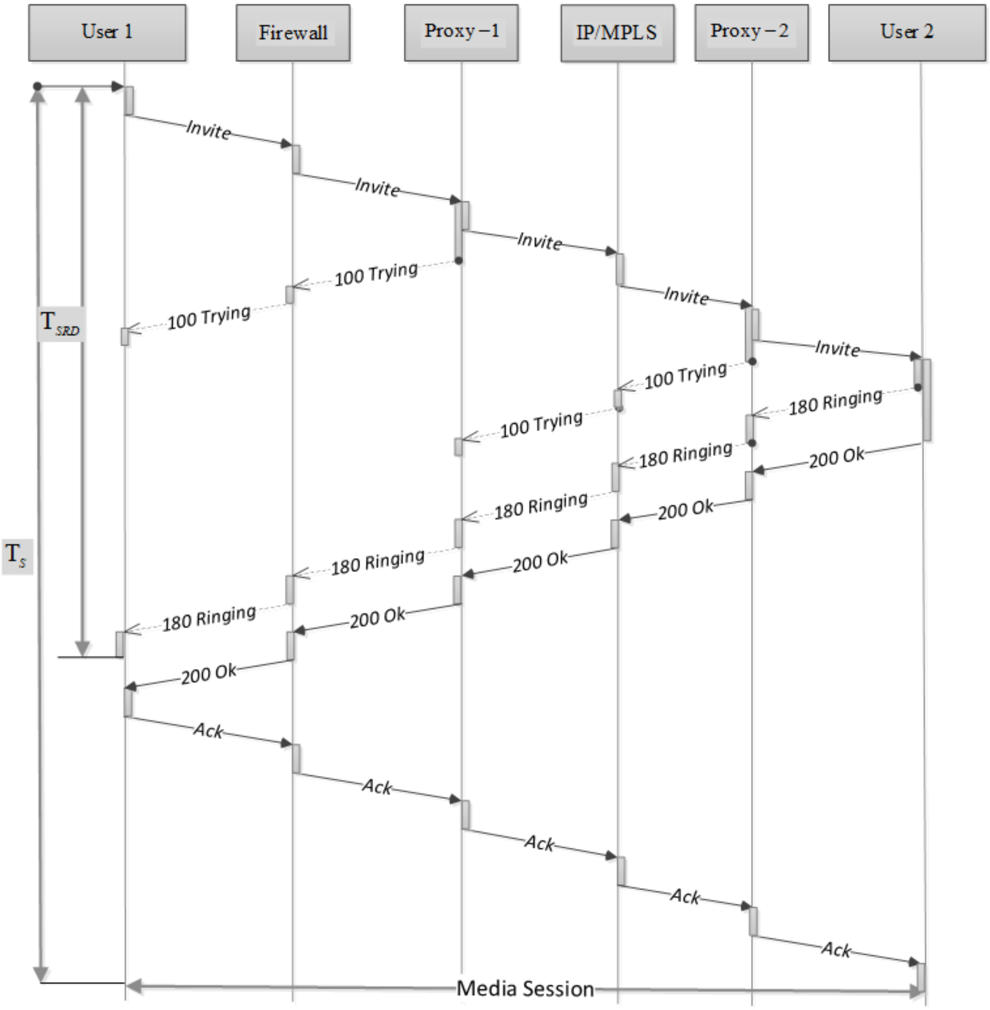


Figure 2. Message sequence when the SIP session is initiated

The residence time in the 2nd block will be equal to the time of the signal message filtration by the firewall:

$$T_F = \frac{1}{\mu_2 - 5\lambda_0}. \tag{4}$$

Using formulas (2)–(4), we determine the indicators of the firewall influence on the session initiation time and the session request delay:

$$N_{T_F\_T_{SRD}} = \frac{\dfrac{2}{\mu_2 - 5\lambda_0} \times 100\%}{2\mu_1^{-1} + \dfrac{2}{\mu_2 - 5\lambda_0} + \dfrac{2}{\mu_3 - 5\lambda_0} + 2\mu_4^{-1} + \dfrac{2}{\mu_5 - 4\lambda_0} + \mu_6^{-1}}; \tag{5}$$

$$N_{T_F\_T_S} = \frac{\dfrac{3}{\mu_2 - 5\lambda_0} \times 100\%}{2\mu_1^{-1} + \dfrac{3}{\mu_2 - 5\lambda_0} + \dfrac{3}{\mu_3 - 5\lambda_0} + 3\mu_4^{-1} + \dfrac{3}{\mu_5 - 4\lambda_0} + 2\mu_6^{-1}}. \quad (6)$$

4. Evaluation of the firewall influence on the session initiation time and the session request delay

To evaluate the firewall influence on the session initiation time and the session request delay, the following Cisco equipment has been selected: the Cisco ASA 5500-X firewall with the SSP-10 module, and the Cisco Sun Fire V120 proxy server. The initial data and their designations are given in Table 1.

Table 1

Initial data

Functional element	User 1	Firewall	Proxy-1	IP/MPLS	Proxy-2	User 2
Designation	μ_1^{-1}	μ_2^{-1}	μ_3^{-1}	μ_4^{-1}	μ_5^{-1}	μ_6^{-1}
Service time, msec.	0.1	0.5	0.4	50	0.4	0.1

The results of the evaluation are presented in the form of graphs showing the dependence of the T_S and T_{SRD} times on the intensity of incoming requests (see the figure 3).

The figure 3 shows that the condition for the existence of the stationary mode (1) makes it possible to provide evaluation at the λ_0 intensity values up to 400 requests per second. The T_S and T_{SRD} values obtained in the presence of the firewall meet the requirements of the international standards for the perception quality indicators. The value of the session initiation time T_S is less than 2 seconds [5], [15]–[18]. At the intensity level $\lambda_0 = 380$ requests per second, the average session initiation time is $T_S = 0.2$ [s], and $T_{SRD} = 0.15$.

The evaluation of the indicators of the firewall influence on the session initiation time and the session request delay is presented in the figure 4.

The firewall residence time for signal messages is less than 10% at the intensity level $\lambda_0 = 370$ [requests/sec].

5. Conclusion

A mathematical model for the SIP session initiation with message filtration by the firewall is presented in this paper. The evaluation of the average session initiation time and the average session request delay indicates the advisability of reducing the residence time that requests spent in the firewall, which can lead to the reduction of the values of QoE and QoS indicators.

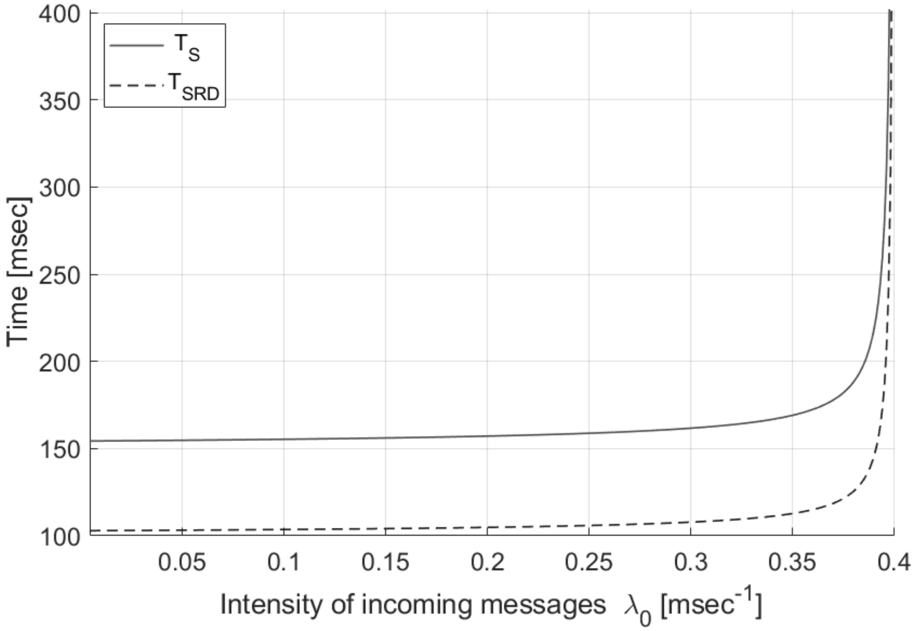


Figure 3. Temporal characteristics when initiating the session with one firewall

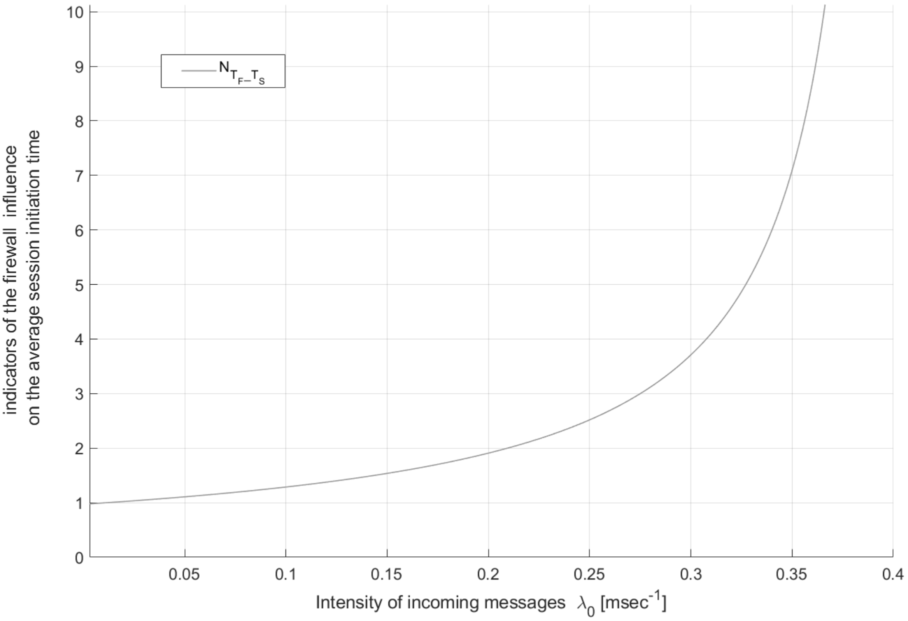


Figure 4. Evaluation of the percentage of the firewall filtering time when initiating the session

References

[1] "Recommendation ITU T G.107. The E model: a computational model for use in transmission planning. Series G: Transmission Systems And Media, Digital Systems And Networks International Telephone Connections And Circuits — Transmission Planning And the E-model," approved in 2015-06-29.

[2] J. Rosenberg, H. Schulzrinne, G. Camarillo, *et al.*, "RFC 3261 SIP: Session Initiation Protocol," 2002.

[3] A. Johnston, S. Donovan, R. Sparks, *et al.*, "RFC 3665 SIP. Session Initiation Protocol (SIP) Basic Call Flow Examples," 2003.

[4] A. B. Goldstein and B. S. Goldstein, *Softswitch.* Saint Petersburg: BHV Publishing House Petersburg, 2006, p. 368.

[5] D. Malas and A. Morton, "RFC 6076. Basic Telephony SIP End to End Performance Metrics," 2011.

[6] K. V. Ivanov and P. I. Tutubalin, *Markov models of protection of automated control systems for special purposes [Markovskie modeli zashhity' avtomatizirovanny'x sistem upravleniya special'nogo naznacheniya].* Kazan: Publishing house of GBU Republican center for monitoring the quality of education Publ., 2012, p. 216, in Russian.

[7] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, "Open, closed and mixed networks of queues with different classes of customers," *Journal of the ACM*, pp. 248–260, 1975. DOI: 10.1145/321879.321887.

[8] K. E. Samouylov, M. V. Luzgachev, and O. N. Plaksina, "Modelling SIP Connections with Open Multiclass Queueing Networks [Razrabotka veroyatnostnoj modeli dlya analiza pokazatelej kachestva protokola iniciirovaniya seansov svyazi]," *Bulletin of Peoples' Friendship University of Russia. Series Mathematics. Information Sciences. Physics*, no. 3, pp. 53–63, 2007, in Russian.

[9] Y. V. Gaidamaka and E. R. Zaripova, "Session Setup Delay Estimation Methods for IMS Based IPTV Services," *Lecture Notes in Computer Science*, vol. 8638, pp. 408–418, 2014. DOI: 10.1007/978-3-319-10353-2_36.

[10] V. M. Vishnevsky, *Polling systems: theory and application in broadband wireless networks [Sistemy pollinga: teoriya i primenenie v shirokopolosnyh besprovodnyh setyah].* Moscow: Technosphere Publishing House, 2007, p. 312, in Russian.

[11] Ali Raad Abdo Mohammed, "Development of a method for evaluating the probabilistic and temporal characteristics of IPTV services when they are controlled by the IMS multimedia subsystem [Razrabotka metoda otsenki veroyatnostno-vremennykh kharakteristik uslug IPTV pri ikh upravlenii mul'timediynoy podsistemoy IMS]," in Russian, Ph.D. dissertation, Moscow technical university of communications and informatics, 2013.

[12] K. E. Samouylov, *Methods of analysis and calculation of ACS networks [Metody analiza i rascheta setey OKS]*. Moscow: Publishing RUDN, 2002, p. 292, in Russian.

[13] I. Buzyukova, Y. Gaidamaka, and G. Yanovsky, "Estimation of QoS parameters in intelligent network," *Lecture Notes in Computer Science*, vol. 5764, pp. 143–153, 2009. DOI: 10.1007/978-3-642-04190-7_14.

[14] K. E. Samouylov, E. S. Sopin, A. V. Chukarin, and A. Y. Botvinko, "Evaluation of the characteristics of signal traffic in the communication network based on the subsystem [Ocenka harakteristik signal'nogo trafika v seti svyazi na baze podsistemy]," *T-Comm — Telecommunications and Transport*, no. 7, pp. 8–13, 2010, in Russian.

[15] "Recommendation ITU T Y.1530. Call processing performance for voice service in hybrid IP networks. Series y: global information infrastructure, internet protocol aspects and next generation networks internet protocol aspects and next-generation networks," approved in 2007-11-13.

[16] "Recommendation ITU T Y.1531. SIP based call processing performance. Series Y: Global Information Infrastructure, Internet Protocol Aspects And Next Generation Networks Internet Protocol Aspects — Quality Of Service And Network Performance," approved in 2007-11-13.

[17] "Recommendation ITU T Y.1541. Network performance objectives for IP based services. Series y: global information infrastructure, internet protocol aspects and next generation networks internet protocol aspects — quality of service and network performance," approved in 2011-12-14.

[18] "DSL Forum, Technical Report-126, Triple-play Services Quality of Experience (QoE) Requirements," 2006.

For citation:

A. Y. Botvinko, K. E. Samouylov, Evaluation of the firewall influence on the session initiation by the SIP multimedia protocol, Discrete and Continuous Models and Applied Computational Science 29 (3) (2021) 221–229. DOI: 10.22363/2658-4670-2021-29-3-221-229.

Information about the authors:

Botvinko, Anatoly Y. — postgraduate of Department of Applied Probability and Informatics (e-mail: botviay@sci.pfu.edu.ru, ORCID: https://orcid.org/0000-0003-1412-981X, Scopus Author ID: 57222085424)

Samouylov, Konstantin E. — Doctor of Technical Sciences, Professor, Head of Department of Applied Probability and Informatics (e-mail: samuylov-ke@rudn.ru, ORCID: https://orcid.org/000-0002-6368-9680, ResearcherID: E-9966-2014, Scopus Author ID: 14009785000)

УДК 519.872
DOI: 10.22363/2658-4670-2021-29-3-221-229

Оценка влияния межсетевого экрана на инициирование сеанса по мультимедийному протоколу SIP

А. Ю. Ботвинко[1], **К. Е. Самуйлов**[1,2]

[1] *Российский университет дружбы народов*
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия
[2] *Федеральный исследовательский центр «Информатика и управление» РАН*
ул. Вавилова, д. 44, корп. 2, Москва, 119333, Россия

Межсетевые экраны — один из основных компонентов обеспечения сетевой безопасности. Используя межсетевые экраны, можно решить такие проблемы, как предотвращение несанкционированного доступа, а также удаление, изменение и/или распространение информации, находящейся под защитой. Процесс фильтрации информационных потоков межсетевым экраном вносит дополнительные задержки по времени, что может привести к нарушению стабильной работы защищаемой автоматизированной системы или недоступности сервисов, предоставляемых системой. Мультимедийные услуги особенно чувствительны к задержкам обслуживания. Основная цель исследования, представленного в статье, — оценить влияние межсетевого экрана на временные задержки в процессе передачи данных в автоматизированной системе с протоколами передачи мультимедийных данных. Оценка обеспечивается методами теории очередей, в то время как сеанс между двумя пользователями инициируется протоколом инициации сеанса (SIP) с фильтрацией сообщений межсетевого экрана. Межсетевой экран — это локальный или функциональный инструмент распределения, который обеспечивает контроль над входящей и/или исходящей информацией в автоматизированной системе (AS) и защиту системы путем фильтрации информации, т.е. гарантирует возможность анализа информации по заданным критериям и принятие решения о её распространении.

Ключевые слова: SIP, межсетевой экран, инициирование сеанса, система очередей, время фильтрации, автоматизированная система

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2021, 29 (3) 230–241
http://journals.rudn.ru/miph

Research article

UDC 519.872
DOI: 10.22363/2658-4670-2021-29-3-230-241

Evaluation of firewall performance when ranging a filtration rule set

Anatoly Y. Botvinko[1], **Konstantin E. Samouylov**[1,2]

[1] *Peoples' Friendship University of Russia (RUDN University)*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[2] *Research Center "Computer Science and Control" of the Russian Academy of Sciences*
44-2, Vavilov St., Moscow, 119333, Russian Federation

(received: August 4, 2021; accepted: September 9, 2021)

This article is a continuation of a number of works devoted to evaluation of probabilistic-temporal characteristics of firewalls when ranging a filtration rule set. This work considers a problem of the decrease in the information flow filtering efficiency. The problem emerged due to the use of a sequential scheme for checking the compliance of packets with the rules, as well as due to heterogeneity and variability of network traffic. The order of rules is non-optimal, and this, in the high-dimensional list, significantly influences the firewall performance and also may cause a considerable time delay and variation in values of packet service time, which is essentially important for the stable functioning of multimedia protocols. One of the ways to prevent decrease in the performance is to range a rule set according to the characteristics of the incoming information flows. In this work, the problems to be solved are: determination and analysis of an average filtering time for the traffic of main transmitting networks; and assessing the effectiveness of ranging the rules. A method for ranging a filtration rule set is proposed, and a queuing system with a complex request service discipline is built. A certain order is used to describe how requests are processed in the system. This order includes the execution of operations with incoming packets and the logical structure of filtration rule set. These are the elements of information flow processing in the firewall. Such level of detailing is not complete, but it is sufficient for creating a model. The QS characteristics are obtained with the help of simulation modelling methods in the Simulink environment of the matrix computing system MATLAB. Based on the analysis of the results obtained, we made conclusions about the possibility of increasing the firewall performance by ranging the filtration rules for those traffic scripts that are close to real ones.

Key words and phrases: firewall, ranging the filtration rules, network traffic, phase service, simulation model, queuing system

1. Introduction

In order to ensure information security of automated systems (AS) that have connections to external untrusted resources, we have to pay attention to the possibility of threats such as violation of confidentiality, integrity and availability of information. A required condition to prevent the threats aimed on violating AS's normal operation is using the firewall technologies [1]–[3].

The main firewall technology is network traffic filtration according to a certain rule set. It is executed at the points of the connection of the AS under protection to external uncontrolled systems and is implemented by using special hardware or software complexes, i.e., firewalls. The firewall filtration rule set is a list of conditions according to which the further transmission of network traffic packets is allowed or denied. The parameters, attributes and characteristics of network traffic flows are usually used to set filtering conditions [4].

The important fact is that the network traffic filtration brings additional time delays during data transmission. High values of the delays during packet filtration can cause packet losses, denials for session initiation and failures in AS's normal work [5], [6].

In works [7]–[13], a great influence of the rule set size and the order of filtration rules in the set on the firewall performance is noted. The influence can be explained by the sequential scheme used to check the packet compliance with the set rules. The maximum decrease in the performance happens while checking the compliance of attributes of packets under filtration with the conditions at the end of the high-dimensional rule set. Defining a rule set that correctly realizes the security policy, but is ineffective in terms of performance, can be considered an error in firewall configuring.

We should also consider that real network traffic has heterogeneity caused by various non-parameterizable factors. This can lead to a decrease in the effectiveness of the static filtration rule set configured initially. One of the ways to prevent the decrease in the performance caused by traffic heterogeneity is to range the rule set according to the incoming traffic characteristics.

Therefore, the task of ranging a rule set in accordance with the characteristics of information flows is not only actual and in demand. This is especially important for the firewalls that ensure information security for the AS with a complex network architecture and large volumes of network traffic. The main goal of this work is to develop a model for evaluating the firewall performance when ranging the filtration rule set.

This paper has the following structure. A method for ranging the filtration rule set is proposed in section 2. In section 3, a model for ranging the rules in the form of a queuing system (QS) with a phase-type service discipline is developed [14]. The results of simulation modelling and firewall performance evaluation for the network traffic script that is close to real are presented in section 4. The Conclusion contains the main aspects of our study.

2. Ranging a filtration rule set for a firewall

By ranging the filtration rule set we mean putting the rules in descending order by their weights in accordance with the evaluation of the characteristics of information flows. We consider that traffic filtration is executed at the network and transmission levels of the standard model for the open system interaction (OSI). According to the generally accepted classification [1]–[3], such firewalls relate to the type of packet filters.

Ranging is executed at discrete moments of time $t_k^- = t_k - 0$ (see the figure 1).

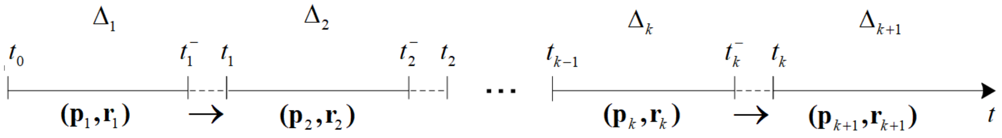

Δ_1
Δ_2
Δ_k
Δ_{k+1}
t_0
t_1^-
t_1
t_2^-
t_2
...
t_{k-1}
t_k^-
t_k
$(\mathbf{p}_1, \mathbf{r}_1)$
→
$(\mathbf{p}_2, \mathbf{r}_2)$
$(\mathbf{p}_k, \mathbf{r}_k)$
→
$(\mathbf{p}_{k+1}, \mathbf{r}_{k+1})$
t

Figure 1. Ranging the filtration rule set

The packets received during the time $\Delta_k = t_k - t_{k-1}$, $k \geqslant 1$ are combined into a package, and a group of q packages generates a redundancy errors in the rule set under data segment. It is assumed that there are no inconsistency and consideration [7]–[10]. We also consider that there is a certain minimum number of rules M under which ranging can provide a significant effect. The logical structure of the filtration rules is a linear list of conditions.

Let us introduce the following designations: M is the number of filtration rules in the set r_i^k — the rule in i-th position in the k-th set of filtration rules on the interval $[t_{k-1}, t_k)$, $\mathbf{r}_k = (r_1^k, \dots, r_M^k)$ k-th set of filtration rules on the interval $[t_{k-1}, t_k)$, p_i^k — the weight of r_i^k rule, $\mathbf{p}_k = (p_1^k, \dots, p_M^k)$ — the rule weight vector on the interval $[t_{k-1}, t_k)$.

In this work, a value of the average number of packets, the attributes of which match the conditions of the r_i^k rule on the interval Δ_k is used as a weight p_i^{k+1}.

The nonparametric method of local approximation (MLA) is used to evaluate the average number of requests [15]–[18]. The same method is used for the analysis of other characteristics investigated in this work.

A method for ranging the rules is proposed in the next section of this paper.

3. The model for ranging the filtration rule set

The complexity and variety of the firewall functioning do not allow to create a model reflecting all the regularities and features that are characteristic for various manufacturers, such as Cisco Systems, Juniper Networks, etc. Therefore, the model describes only the main regularities and factors of the firewall functioning that are of interest for our tasks.

For all firewall types in the process of network traffic filtration, the following stages can be distinguished [10]:

- — initial packet processing, i.e., operations with a packet when it enters the receiving path;
- — checking the filtration rule set;
- — completion of packet processing, i.e., operations with a packet when it is transmitted to the output path and then to the physical medium.

During the initial processing of the packet received the firewall network interface controller (NIC) decodes the sequence of electrical or optical signals, checks the correctness of information delivered and writes the packet into the NIC input buffer memory. Then the packet is transmitted to a program buffer located in RAM for operations executed by the central process.

As a next step, if computing resources are available, the filtration of the incoming packet is executed in accordance with the filtration rule set. The compliance of the received packet parameters with the filtration rules is checked in sequence. Only one packet can be checked at a time. The other packets received are kept in the buffer. Their service is executed according to the order of their entrance to the buffer (FCFS, First-Come, First-Served).

If the packet parameters match the filtration rules, the firewall transmits the packet to the NIC output buffer. If the packet parameters do not match the permissive rules, the firewall rejects the packet. The packet processing is considered complete when it is encoded and transmitted to the physical medium.

Let us present the firewall model as a queuing system (QS) with a $B_k(t)$ distribution function (DF) for the request service duration, which depends on the order of the filtration rules on time interval $[t_{k-1}, t_k)$. A request flow $\Lambda(t)$, corresponding to the packet flow incoming the firewall, enters the QS. We consider the incoming packets as the service requests for the QS.

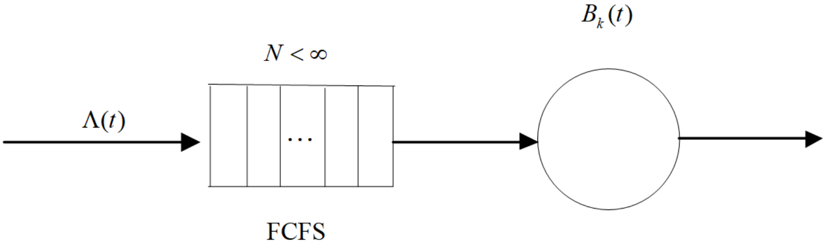


Figure 2. The scheme of the firewall QS with a complex request service discipline

The $B_k(t) = B(\mathbf{r}_k, \mu_0, \mu, t)$ distribution function (DF) is a function of phase type, its parameters are shown in the figure 3, from which it is clear that the $B_k(t)$ DF corresponds to the Cox distribution [19].

The request service time at zero phase corresponds to the total time of packet initial processing and the time of transmission along the output path. The request service time at the m-th phase $m > 1$ corresponds to the time of packet filtration the by the m rule. It is assumed that the filtration time for each rule is the same and equal to τ.

The scheme of the request service process in the firewall model is presented in the figure 3.

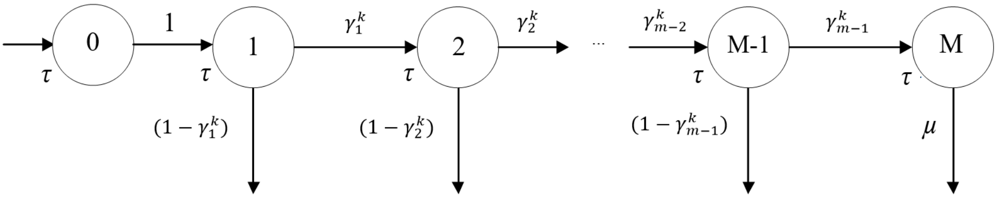


Figure 3. The request service process

The $1 - \gamma_i^k$ value corresponds to the probability of completing the request service at the i-th phase. That is the case when the packet attributes do not correspond to the r_i^k rule. Therefore, the DF of the request service time in the QS on the interval $[t_{k-1}, t_k)$ is as follows:

$$B_k(t) = \gamma_1^k E(1, \mu_0) + \sum_{i=1}^{M} \gamma_1^k E(i, \mu), \tag{1}$$

where $E(1, \mu_0)$ is the Erlang distribution of the i-th order.

The task of analyzing the QS (shown in Fig. 2) characteristics can be solved with the help of the simulation modelling method, the results of which are presented in the next section. It should be noted that in case of a Poisson incoming flow and exponential filtering time, the QS has an analytical solution [14].

4. Evaluation of the firewall performance when ranging the rules

Firewall is a network node processing large volume of incoming and outgoing traffic. Therefore, the average packet filtering time is usually used as the major performance indicator [3], [7]. In this work, to evaluate the firewall performance, we use ΔU_S, i.e., a value equal to the difference between U_1 — the average filtering time in the first data segment (without rule ranging) and U_S — the average filtering time in the S-th data segment (after the rules ranged).

The initial data used for the implementing the simulation model of the process of network traffic filtration are shown in the table 1.

Table 1

Initial data

[rules]	μ_0^{-1} [ms]	μ^{-1} [ms]	$\Delta_k, k = 1, ..., 25$ [ms]	q [packages]	s [segments]
100	$2.7 \cdot 10^{-3}$	$5 \cdot 10^{-5}$	1000	5	5

The number of packet types is M. The values of request service intensities — μ_0 and μ — have been taken from the work [10], which is about the analysis of the firewall performance under the Poisson incoming flow of requests.

To provide the numerical analysis of the QS (see the figure 1), a simulation model (SM) is built in the Simulink simulation environment of the MATLAB matrix computing system with the use of SimEvents discrete state library. The scheme of the model is presented in the figure 4.

The request flow in the SM is determined in the Traffic Generation subsystem. A request collector is realized by the FIFO Queue block, and the request service process is executed by the Single Server blocks (the QS service device) and the `function_f` subsystem (the calculation of the request service time in accordance with the rule set and request type).

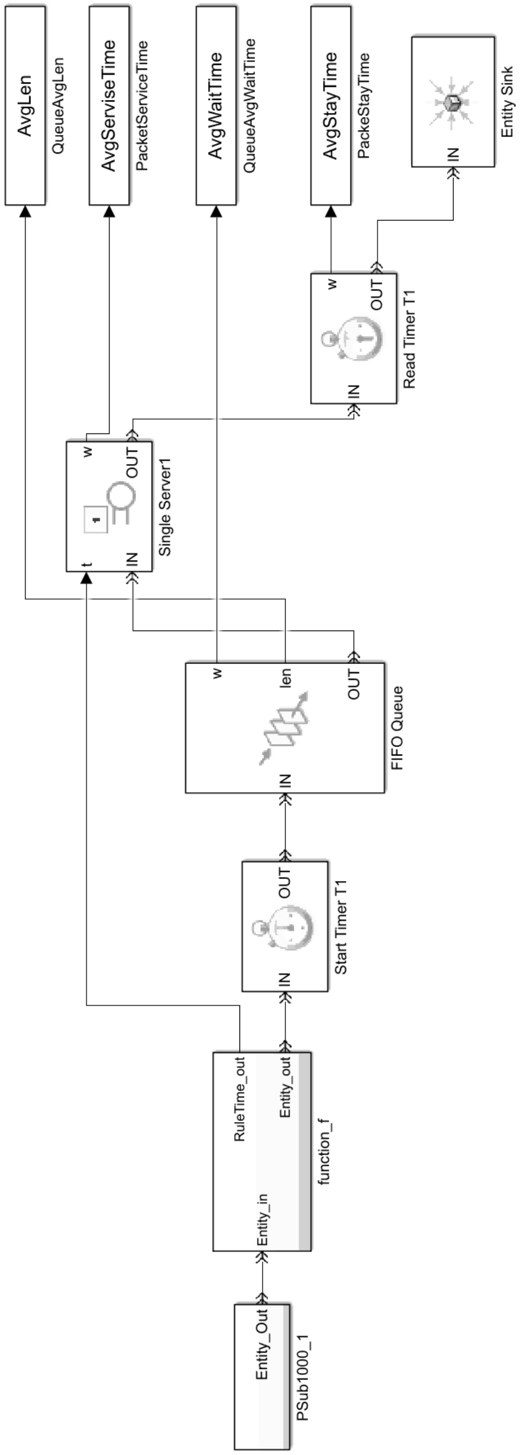


Figure 4. The scheme of the simulation model in the Simulink environment

The statistical data accumulation for evaluating the performance indicators is executed with the help of the statistics collection options of the SimEvents blocks and data recording structures such as `PacketServiceTime`, `QueueAvgWaitTime`, `PackeStayTime`, `QueueAvgLen`. The following times are fixed at this stage: the packet service time, the waiting time for the packet in the queue, the average time of the packet residence in the system, and the average length of the packet queue.

To define the incoming flow of requests, data from the WIDE academic core network in Japan have been used. Traffic records are contained in the MAWI Group Traffic Archive traffic repository by 01/10/2019. For each packet type, using the Wireshark tool for network traffic capture and analysis, the values of the time intervals between packets for the TCP, UDP and ICMP protocols have been extracted. The data massive obtained has been exported to MATLAB to set the intervals between the moments of request generation in the Traffic Generation subsystem using Time-Based Entity Generator blocks. The request types corresponding to the traffic packet types are determined in the Traffic Generation subsystem by the SetPacketAtt blocks. An example of the request flow obtained for packets of r^1_{81} and r^1_{29} types is presented in figures 5–6.

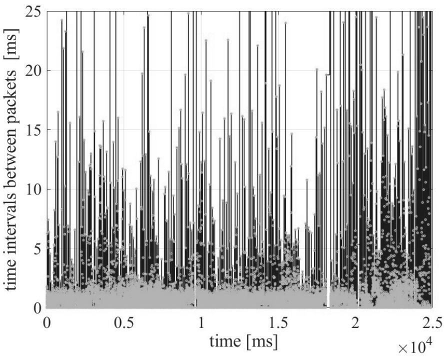


Figure 5. The packet flow of 81$^{\text{st}}$ type

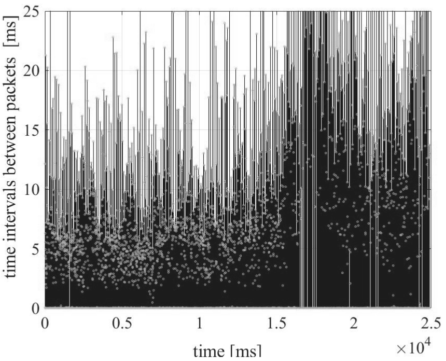


Figure 6. The packet flow of 29$^{\text{th}}$ type

The following actions are implemented in M-files of the MATLAB system: determination of the initial data for simulation modelling (see the table 1) and the initial rule set, calculation of the performance indicators, execution of functions for calculating weight, rule set ranging and other algorithms and SM variables.

The process of ranging the r^1_{81} and r^1_{29} filtration rules in accordance with the evaluation of the information flow characteristics is illustrated by figures 7–8. The figures show that:

— the r^1_{81} rule, when ranging, takes the 7th place in the set (average). This can be explained by the short time interval between the packet income (4 ms) and the small value of the time dispersion between the income of the packets;

— the r^1_{29} rule is characterized by moving to the middle of the set. It happens due to the increase in the time interval between the income of the packets.

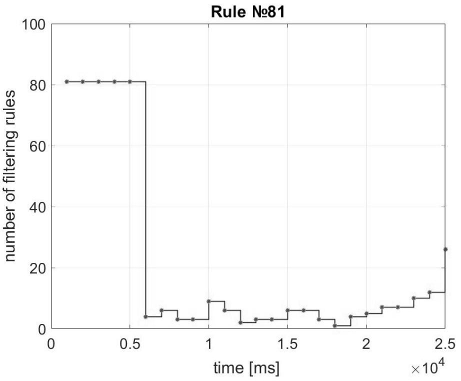


Figure 7. Ranging the r^1_{81} rule

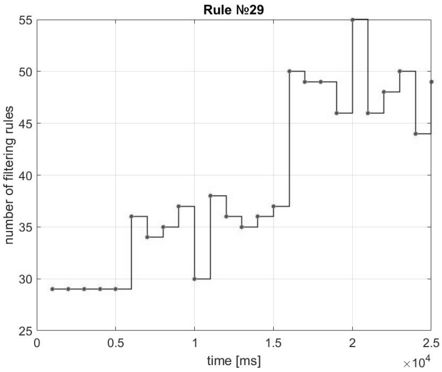


Figure 8. Ranging the r^1_{29} rule

Simulation modelling has demonstrated that the average packet filtration time for all time intervals $[t_{k-1}, t_k) \in T$, $k > 5$ on which ranging has been executed, has a decrease compared to the average time on the intervals $[t_{k-1}, t_k) \in T$, $k = 1, \dots, 5$.

For the first interval $[t_0, t_1)$, where there is no set ranging, and for the last interval $[t_{24}, t_{25})$, where the set is ranged, we can present the graph of the average packet filtration time (see the figure 9).

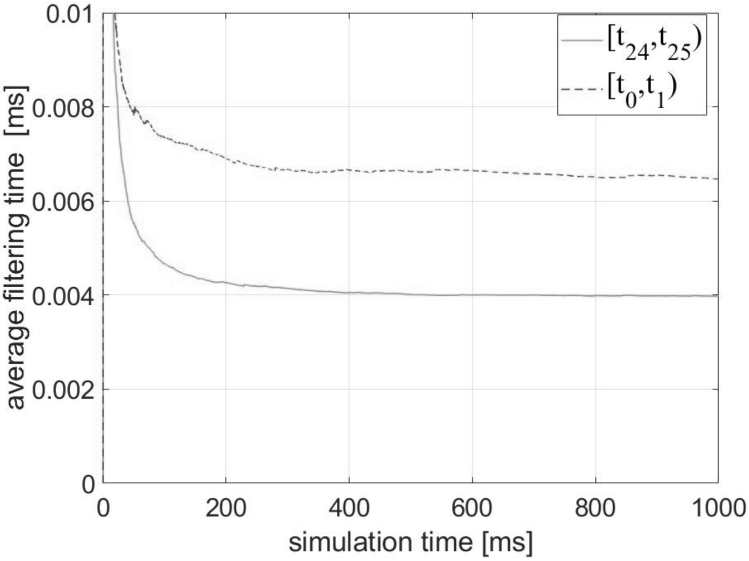


Figure 9. Average packet filtration time for first and the last intervals

As can be seen from the figure, the value of the average packet filtration time on the interval $[t_0, t_1)$ is larger than the average time on the interval $[t_{24}, t_{25})$ by about 2.5 [s]. The results of the firewall performance evaluation for all segments obtained during the simulation modelling are presented in the table 2.

Table 2

The firewall performance

s	U_S [s]	ΔU_S [s]	ΔU_S [%]
1	6.233	-	-
2	4.937	1.296	20.785
3	4.660	1.573	25.229
4	4.989	1.244	19.960
5	4.406	1.827	29.304

5. Conclusion

The created QS with a complex request service discipline and the simulation methods allowed us to obtain the firewall performance estimates when ranging a rule set. These estimates demonstrate that, for the traffic of the main transmission networks, ranging has increased the firewall performance by 20–29% compared to traffic filtering without ranging. So, the results obtained indicate the possibility of increasing the firewall performance for traffic scripts that are close to real ones. These results also confirm the assumptions made in work [20] about the advisability of ranging.

The authors plan to study the influence of the ranging interval and MLA parameters on the firewall performance in further works. They also plan to develop criteria for the need of re-ranging the set depending on changes in the firewall performance indicators, as well as recommendations for ranging the filtration rule sets.

References

[1] S. V. Lebed, *Firewall protection. Theory and practice of external perimeter protection [Mezhsetevoye ekranirovaniye. Teoriya i praktika zashchity vneshnego perimetra]*. Moscow: BMSTU, Bauman Moscow State Technical University Publ., 2002, p. 304, in Russian.

[2] O. R. Laponina, *The foundation of network security [Osnovy setevoy bezopasnosti]*. Moscow: Publishing house of the national Open University «INTUIT», 2014, p. 377, in Russian.

[3] K. V. Ivanov and P. I. Tutubalin, *Markov models of protection of automated control systems for special purposes [Markovskie modeli zashhity' avtomatizirovanny'x sistem upravleniya special'nogo naznacheniya]*. Kazan: Publishing house of GBU Republican center for monitoring the quality of education Publ., 2012, p. 216, in Russian.

[4] "Governing document. Computer aids. Firewall. Protection against unauthorized access to information. Indicators of security against unauthorized access to information [Rukovodyashhij dokument. Sredstva vy'chislitel'noj texniki. Mezhsetevy'e e'krany'. Zashhita ot nesankcionirovannogo dostupa k informacii. Pokazateli zashhishhennosti ot nesankcionirovannogo dostupa k informacii] approved by the decision of the Chairman of the State Technical Commission under the President of the Russian Federation dated July 25, 1997," in Russian.

[5] H. Hamed, A. El-Atawy, and E. Al-Shaer, "On dynamic optimization of packet matching in high-speed firewalls," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 10, pp. 1817–1830, 2006. DOI: `10.1109/JSAC.2006.877140`.

[6] R. Mohan, A. Yazidi, B. Feng, and J. Oommen, "On optimizing firewall performance in dynamic networks by invoking a novel swapping window-based paradigm," *International Journal of Communication Systems*, vol. 31, no. 15, e3773, 2018. DOI: `10.1002/dac.3773`.

[7] E. Al Shaer, *Automated firewall analytics: Design, configuration and optimization*. Springer International Publishing, 2014, p. 132. DOI: `10.1007/978-3-319-10371-6`.

[8] R. Mohan, A. Yazidi, B. Feng, and B. J. Oommen, "Dynamic ordering of firewall rules using a novel swapping window-based paradigm," in *Proceedings 6th International Conference on Communication and Network, ICCNS 2016*, Singapore: ACM Proceedings, 2016, pp. 11–20. DOI: `10.1145/3017971.3017975`.

[9] Z. Trabelsi, S. Zeidan, M. M. Masud, and K. Ghoudi, "Statistical dynamic splay tree filters towards multilevel firewall packet filtering enhancement," *Computers & Security*, vol. 53, pp. 109–131, 2015. DOI: `10.1016/j.cose.2015.05.010`.

[10] K. Salah, K. Elbadawi, and R. Boutaba, "Performance modeling and analysis of network firewalls," *IEEE Transactions on Network and Service Management*, vol. 9, no. 1, pp. 12–21, 2012. DOI: `10.1109/TNSM.2011.122011.110151`.

[11] C. Diekmann, L. Hupel, J. Michaelis, M. Haslbeck, and G. Carle, "Verified iptables firewall analysis and verification," *Journal of Automated Reasoning*, vol. 61, no. 1–4, pp. 191–242, 2018. DOI: `10.1007/s10817-017-9445-1`.

[12] S. Khummanee, "The semantics loss tracker of firewall rules," *Advances in Intelligent Systems and Computing*, vol. 769, pp. 220–231, 2018. DOI: `10.1007/978-3-319-93692-5_22`.

[13] V. Clincy and H. Shahriar, "Detection of anomaly in firewall rule-sets," *Advances in Intelligent Systems and Computing*, vol. 842, pp. 422–431, 2018. DOI: `10.1007/978-3-319-98776-7_46`.

[14] P. P. Bocharov and A. V. Pechenkin, *Queuing theory [Teoriya massovogo obsluzhivaniya]*. Moscow: Publishing RUDN, 1995, p. 529, in Russian.

[15] V. Y. Katkovnik, *Non-parametric data identification and smoothing: local approximation method [Neparametricheskaya identifikatsiya i sglazhivaniye dannykh: metod lokal'noy approksimatsii]*. Moscow: The science. Main editorial office of physical and mathematical literature Publ., 1985, in Russian.

[16] J. M. Bravo, T. Alamo, M. Vasallo, and M. E. Gegúndez, "A general framework for predictors based on bounding techniques and local approximation," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3430–3435, 2017. DOI: 10.1109/TAC.2016.2612538.

[17] H. Al-Shuka, "On local approximation-based adaptive control with applications to robotic manipulators and biped robots," *International Journal of Dynamics and Control*, vol. 6, no. 1, pp. 339–353, 2018. DOI: 10.1007/s40435-016-0302-6.

[18] D. E. Plotnikov, T. S. Miklashevich, and S. A. Bartalev, "Using local polynomial approximation within moving window for remote sensing data time-series smoothing and data gaps recovery [Vosstanovleniye vremennykh ryadov dannykh distantsionnykh izmereniy metodom polinomialnoy approksimatsii v skolzyashchem okne peremennogo razmera]," *Modern problems of remote sensing of the Earth from space of the Russian Academy of Sciences*, vol. 11, no. 2, pp. 103–110, 2014, in Russian.

[19] D. R. Cox, "A use of complex probabilities in the theory of stochastic processes," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 2, pp. 313–319, 1955. DOI: 10.1017/S0305004100030231.

[20] A. Y. Botvinko and K. E. Samouylov, "Adaptive ranking of the firewall rule set using local approximation [Adaptivnoye ranzhirovaniye nabora pravil mezhsetevogo ekrana metodom lokal'noy approksimatsii]," in *Distributed Computer and Communication Networks: Control, Computation, Communications*, in Russian, 2018, pp. 334–341.

For citation:

A. Y. Botvinko, K. E. Samouylov, Evaluation of firewall performance when ranging a filtration rule set, Discrete and Continuous Models and Applied Computational Science 29 (3) (2021) 230–241. DOI: 10.22363/2658-4670-2021-29-3-230-241.

Information about the authors:

Botvinko, Anatoly Y. — postgraduate of Department of Applied Probability and Informatics (e-mail: botviay@sci.pfu.edu.ru, ORCID: https://orcid.org/0000-0003-1412-981X, Scopus Author ID: 57222085424)

Samouylov, Konstantin E. — Doctor of Technical Sciences, Professor, Head of Department of Applied Probability and Informatics (e-mail: samuylov-ke@rudn.ru, ORCID: https://orcid.org/000-0002-6368-9680, ResearcherID: E-9966-2014, Scopus Author ID: 14009785000)

УДК 519.872
DOI: 10.22363/2658-4670-2021-29-3-230-241

Оценка производительности межсетевого экрана при ранжировании набора правил фильтрации

А. Ю. Ботвинко[1], **К. Е. Самуйлов**[1, 2]

[1] *Российский университет дружбы народов*
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия
[2] *Федеральный исследовательский центр «Информатика и управление» РАН*
ул. Вавилова, д. 44, корп. 2, Москва, 119333, Россия

Данная статья является продолжением ряда работ, посвящённых оценке вероятностно-временных характеристик межсетевых экранов при ранжировании набора правил фильтрации. В публикации рассматривается проблема снижения эффективности фильтрации информационных потоков. Проблема возникла из-за использования последовательной схемы проверки соответствия пакетов правилам, а также из-за неоднородности и изменчивости сетевого трафика. Порядок правил неоптимален, и это в многомерном списке существенно влияет на производительность межсетевого экрана, а также может вызывать значительную временную задержку и вариации в значениях времени обслуживания пакетов, что существенно важно для стабильной работы мультимедийных протоколов. Один из способов предотвратить снижение производительности — это ранжировать набор правил в соответствии с характеристиками входящих информационных потоков. В исследовании решаются следующие задачи: определение и анализ среднего времени фильтрации трафика основных передающих сетей; оценка эффективности ранжирования правил. Предложен метод ранжирования набора правил фильтрации и построена система массового обслуживания со сложной дисциплиной обслуживания запросов. Определённый порядок используется для описания того, как запросы обрабатываются в системе, и включает в себя выполнение операций с входящими пакетами и логическую структуру набора правил фильтрации. Таковы элементы обработки информационного потока в межсетевом экране. Подобный уровень детализации не полный, но его достаточно для создания модели. Характеристики СМО получены с помощью методов имитационного моделирования в среде Simulink матричной вычислительной системы MATLAB. На основании анализа полученных результатов были сделаны выводы о возможности повышения производительности межсетевого экрана за счёт ранжирования правил фильтрации для тех скриптов трафика, которые близки к реальным.

Ключевые слова: межсетевой экран, ранжирование правил фильтрации, сетевой трафик, фазовое обслуживание, имитационная модель, система массового обслуживания

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2021, 29 (3) 242–250
http://journals.rudn.ru/miph

Research article

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2021-29-3-242-250

Towards the analysis of the performance measures of heterogeneous networks by means of two-phase queuing systems

Tatiana V. Rykova

Peoples' Friendship University of Russia (RUDN University)
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation

(received: August 16, 2021; accepted: September 9, 2021)

Due to a multistage nature of transmission processes in heterogeneous 4G, 5G mobile networks, multiphase queuing systems become one of the most suitable ways for the resource allocation algorithms analysis and network investigation. In this paper, a few scientific papers that approached heterogeneous networks modelling by means of multiphase queuing systems are reviewed, mentioning the difficulties that arise with this type of analytical analysis. Moreover, several previously investigated models are introduced briefly as an example of two-phase systems of finite capacity and a special structure in discrete time that can be used for analysing resource allocation schemes based on the main performance measures obtained for wireless heterogeneous networks. One of the model presents a two-phase tandem queue with a group arrival flow of requests and a second phase of the complex structure that consists of parallel finite queues. The second model is a two-phase tandem queue with Markov modulated geometric arrival and service processes at the first phase and exhaustive service process at the second phase, which solves a cross-layer adaption problem in a heterogeneous network.

Key words and phrases: two-phase model, queuing system, Markov chain, resource allocation, heterogeneous networks

1. Introduction

The Fifth Generation (5G) mobile networks are characterized by advanced algorithms for time-frequency resource allocation schemes in a heterogeneous cell between Base Station (BS) and User Equipment (UE) [1], [2]. Due to a multistage nature of transmission processes in the heterogeneous environment, multiphase queuing systems become one of the most suitable ways for the resource allocation algorithms analysis and network investigation. In [3], researchers have proposed to use single-phase queuing systems for modelling local networks, by giving the necessary physical meaning to the stages of service process using a phase-type service distribution. However, in the case of the Next Generation Mobile Networks (NGMN) the given assumptions are not able to take into account the complex structure of a network with intermediate storage of transmitted information. A large number of publications [4]–[6] are devoted to the analysis of multiphase queuing systems that consider various variants of structural parameters: capacitance of the buffers at the phases, the number of servers at phases, an ordinary or non-ordinary arrival flow, blocking of service at a phase or loss of a request given that the buffer of the next phase is fully occupied, the possibility of the retransmission at the phase or in the system in general, and various arrival and service distributions of requests. In the given publications, the number of phases is usually limited to two, and they are considered mainly in continuous time.

Only a few works approached to investigate heterogeneous networks by means of multiphase (two-phase) queuing systems in discrete time, see, for example, [7], [8]. However, the models in [7], [8] cannot be used because they do not take into account the complex phase structure when modelling transmission processes in a cell and, therefore, do not fully correspond to solving a resource allocation problem in a context of a NGMN cell. It should be noted that most of the foreign publications when using “discrete” and “tandem queue” terms in their papers cover, in fact, mean cyclical service systems in discrete time, but not multiphase systems.

In most of the cases, the number of phases in a multiphase queuing system that is taken as a mathematical model for analysis of the performance measures in a NGMN cell should be taken equal to two. This is due to the fact that each phase itself is a structurally complex queuing system with complex rules of functioning, and a further increase of phases severely complicates formalization of the entire system, leads to multidimensional processes that describe its behaviour and a difficult practical use. The analysis in this case becomes extremely bulky with high risks of obtaining inaccurate results. The decomposition of such a system with the analysis of individual phases or groups of phases is most often not applicable due to the significant mutual influence of phases, in contrast to almost completely decomposable systems [9], [10], and can lead in most cases to significant modelling errors. Cases of independence of the functioning of a phase from the previous phase and, accordingly, an admissible decomposition are rare and arise when conditions [11], are met, for example, when using exponential distributions and buffers of unlimited capacity [12], or under assumptions about specific conditions for the functioning of phases [13]. Summarizing all of that mentioned above, in this paper we briefly overview several two-phase systems of finite capacity in discrete time of a special structure that can be used for analysing resource allocation schemes based on the main performance measures obtained for wireless heterogeneous networks.

2. Two-phase model in discrete time for resource allocation analysis in heterogeneous networks

Heterogeneous networks with the utilization of lower power levels Relay Nodes (RN) improve the capacity of the system, coverage due to the availability of the alternative paths to users, located in shadow areas, and lower deployments costs. Moreover, relay nodes are characterized by wireless backbone access. However, to achieve its potential, the heterogeneous networks are to utilize an efficient cooperative resource allocation procedure on various paths, e.g. from the base station (gNB, gNodeB) to the RN, and from the RN to the User Equipment (UE), in order to avoid data shortage or overflow of the data at relay nodes. An analytical model of heterogeneous network in terms of a two-phase model in discrete time is further introduced, that presents an efficient tool to study resource allocation procedures by means of the found stationary probability distribution and derived performance measures.

2.1. Model's description

Let us consider downlink transmission in a heterogeneous network with K RNs and a single gNB with a subframe that is divided into S channels, which are distributed between the heterogeneous nodes in a centralized manner. The figure 1 demonstrates the structure of the given model, and the main parameters are shown in the table 1. As can be seen from the table, the arrival rate follows a $(K+1)$-dimensional group geometrical distribution. In its turn, the service time in both phases is selected to follow deterministic law equal to one time slot, which corresponds to the transmission of one packet. After the request is being serviced it occupies one space in the buffer.

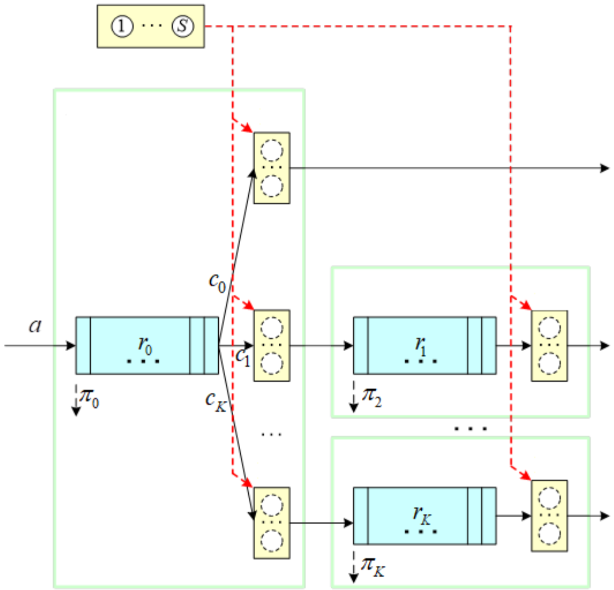


Figure 1. Structural representation of the two-phase model with S channels

The functioning of the given model is described by the homogeneous Markov chain ξ_n at time moments $nh+0, n \geqslant 0$, with the following state space:

$$X = \left\{ \vec{x} = (x_0, x_1, \dots, x_K)^T : \; x_k = 0, \dots, r_k, \; k = 0, \dots, K \right\},$$

$$|X| = \prod_{k=0}^{K} (r_k + 1),$$

Table 1

Definition of the main parameters used in the model.

Parameter	Description
0-request	0-type request to be send to the UE in the coverage area of gNB
k-request	k-type request to be send to the UE in the coverage area of k-RN, $k = 1, 2, \dots, K$
r_0	buffer capacity of the gNB, $r_0 < \infty$
r_k	buffer capacity of the k-RN, $r_k < \infty$, $k = 1, 2, \dots, K$
$h = 1$	constant length of a time slot, in which the system functioning is measured and is equal to LTE downlink data subframe
a	arrival rate that follows a group geometric distribution Geom^G
c_k	probability of a request from the arrival group belonging to type k, k-request, $k = 0, 1 \dots, K$
$c_.$	full sum of the variable c_k, $k = 0, 1 \dots, K$

In the table 1, x_k is a number of k-requests stored in the buffer of corresponding heterogeneous node: gNB or k-RN. Please refer to [14] for more details on derivation of stationary probability distribution and the main performance measures. One of the advantages of the given analytical model is the ability to study various resource allocation procedures by utilizing the following vector in the balance equations:

$$s^n = (s_0^n, s_1^n, \dots, s_K^n)^T = (f_0^n(\vec{x}), f_1^n(\vec{x}), \dots, f_K^n(\vec{x})),$$

where $f^n(\vec{x})$ is a function that introduces the resource allocation strategy. The definition of the different resource allocation procedures and experimental analysis can be found in [14]. All in all, the given model allows analysing various resource assignment schemes, including dynamic schemes, e.g. proportional fair [15] and with fixed allocation.

3. Two-phase model in discrete time for cross-layer optimization in heterogeneous networks

Video transmission comes along with huge demands on resources and low delay, which can be provided by means of Cross-Layer Adaptation (CLA) principle. The given principle is responsible for optimizing the selected metric by adapting the parameters of different layers of open systems interconnection model. The common assumption is to locate CLA mechanism at the gNB, which brings certain shortcomings in terms of the achieved performance. In this paper, we introduce a two-phase analytical model in discrete time that evaluates the behaviour of a downlink video transmission system using CLA principle. We assume Dynamic Adaptive Streaming over HTTP (DASH) in our modelling, the details of which can be found in [16]. The model introduced below covers both the video delivery from gNB to UE at the first phase and video processing at UE at the second phase. Moreover, the CLA is achieved by varying the arrival rate based on the received DASH message, and service probability based on the Channel Quality Indicator (CQI) sent from the UE.

3.1. Model's description

We assume a DASH-based video transmission from the gNB to a single UE in a heterogeneous network. The figure 2 demonstrates the structure of the given model, and the main parameters are shown in the table 2. The functioning of the given model is described by the homogeneous Markov chain ξ_n at time moments $nh + 0, n \geqslant 0$, with the following state space:

$$X = \{\vec{x} = (q_1, q_2, s) : q_1 = 0, 1, \dots, r_1,\ q_2 = 0, 1, \dots, r_2 - 1,\ s = 1, 2, \dots, S\},$$

where q_1 and q_2 are the numbers of requests at the first and second phase, respectively, and s is a value of the CQI in the current state.

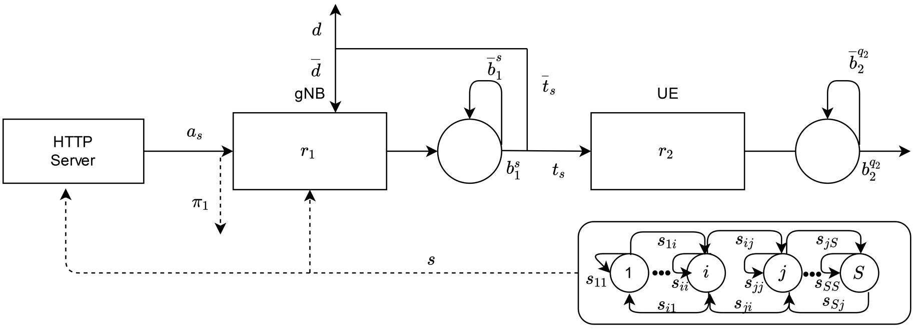


Figure 2. Structural representation of the two-phase model

It should be noted that the stationary probability distribution can be found in a matrix recursive form [13]. However, due to the fact that the functioning of the first phase along with variation of the CQI are independent from the second phase, conditions [11] fulfilled, it is possible to decompose the system to analyse the systems separately, which allows reducing the computational complexity. The conducted experimental analysis of the main performance measures derived from the stationary distribution can be found in [13]. The given two-phase model presents an efficient tool that covers video transmission process from gNB to UE at the first phase and the video decoding process at the UE at the second phase. It takes into account CLA principle, along with the losses due to fading and retransmission.

Table 2

Definition of the main parameters used in the model

Parameter	Description
r_1	buffer capacity of the gNB, $r_1 < \infty$
r_2	buffer capacity of the UE, $r_2 < \infty$
h	constant length of a time slot, in which the system functioning is measured and is equal to LTE downlink data subframe, 1 ms
s	CQI report, which is available both at gNB and UE every time slot $s = 1, 2 \ldots, S$, where S is an overall number of its values
s_{ij}	transition probability of s from state i to state j
a_s	arrival rate that follows geometric distribution
b_1^s	service time at the first phase that follows geometric distribution
$\bar{t}_s\bar{d}$	retransmission probability due to wireless channel errors
$\bar{t}_s d$	probability that the packet lifetime is expired and cannot be used for the video playback
$b_2^{q_2}$	service time at the second phase that follows geometric distribution in exhaustive manner

4. Conclusions

This paper reviews a few scientific papers that approached heterogeneous networks modelling by means of multiphase queuing systems and stressed the difficulties that arise with this type of analytical modelling. Two efficient two-phase models were briefly introduced that can be used for analysing resource allocation schemes based on the main performance measures obtained for wireless heterogeneous networks. One of the model presents a two-phase tandem queue with a group arrival flow of requests and a second phase of the complex structure that consists of parallel finite queues. The second model is a two-phase tandem queue with Markov modulated geometric arrival and service processes at the first phase and exhaustive service process at the second phase, which solves a cross-layer adaption problem in a heterogeneous network.

References

[1] "ITU-R M.2134. Requirements Related to Technical Performance for IMT-Advanced Radio Interface(s)," 2018.

[2] E. Medvedeva, A. Gorbunova, Y. Gaidamaka, and K. Samuylov, "A Discrete Queueing Model for Performance Analysis of Scheduling Schemes in Multi-User MIMO Systems," IEEE, 2019, pp. 1–5. DOI: `10.1109/ICUMT48472.2019.8970876`.

[3] G. P. Basharin and V. A. Efimushkin, *Graph-matrix models of local area networks [Grafomatrichnyye modeli lokal'noy seti].* Moscow: UDN, 1986, in Russian.

[4] T. V. Efimushkina, "Performance evaluation of a tandem queue with common for phases servers," in *18-th International Conference on Distributed Computer and Communication Networks (DCCN-2015): Control, Computation, Communications*, ICS RAS, Moscow, Russia: Technosphere, 2015, pp. 44–51.

[5] V. Klimenok and A. Dudin, "Dual tandem queueing system with multiserver stations and retrials," in *International Conference on Distributed computer and communication networks: control, computation, communications (DCCN-2013)*, ICS RAS, Moscow, Russia, 7–10 October 2013: Technosphere, 2013, pp. 394–401.

[6] P. P. Bocharov, A. V. Pechinkin, and S. Sanchez, "Stationary state probabilities of a two-phase queueing system with a markov arrival process and internal losses," in *Proc. of the Fourth Int. Workshop on Queueing Networks with Finite Capacity*, Ilkley, 2000, pp. 06/1–10.

[7] G. P. Basharin and V. A. Efimushkin, "Algorithmic Analysis of Structurally Complex Systems of Finite Capacity with a Two-Dimensional State Space," in *Teletraffic Theory Models in Communication Systems and Computer Technology.* Nauka, 1985, pp. 28–41.

[8] E. V. Viskova, *Two-phase queuing system with Markov flow and discrete-time service [Dvukhfaznaya sistema massovogo obsluzhivaniya s markovskimi potokom i obsluzhivaniyem v diskretnom vremeni]*, 3. Information processes, 2005, vol. 5, pp. 247–257, in Russian.

[9] G. P. Basharin, P. P. Bocharov, and Y. A. Kogan, *Analysis of queues in computer networks. Theory and methods of calculation [Analiz ocheredey v vychislitel'nykh setyakh. Teoriya i metody rascheta].* Moscow: Nauka, 1989, p. 336, in Russian.

[10] P. J. Courtois, *Queueing and computer system application.* New York: Academic Press, 1977.

[11] V. A. Naumov, "On the independent operation of subsystems of a complex system," Russian, in *Proceedings of the 3rd All-Union School-Meeting on the Theory of Queuing*, in Russian, Moscow, Russia: MSU, 1976, pp. 169–177.

[12] J. R. Jackson, *Networks of Waiting Lines*, 4. Operations Research, 1957, vol. 5, pp. 518–521.

[13] T. V. Efimushkina, M. Gabbouj, and K. E. Samuylov, "Analytical model in discrete time for cross-layer video communication over LTE," *Automatic Control and Computer Sciences*, vol. 48, no. 6, pp. 345–357, 2014. DOI: `10.3103/S0146411614060029`.

[14] T. V. Efimushkina and K. E. Samuylov, "Analysis of the Resource Distribution Schemes in LTE-Advanced Relay-Enhanced Networks," *Communications in Computer and Information Science*, vol. 279, pp. 43–57, 2014. DOI: 10.1007/978-3-319-05209-0_4.

[15] D. Avidor, S. Mukherjee, J. Ling, and C. Papadias, "On some properties of the proportional fair scheduling policy," in *2004 IEEE 15th International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE Cat. No.04TH8754)*, vol. 2, 2004, pp. 853–858. DOI: 10.1109/PIMRC.2004.1373820.

[16] "ISO/IEC 23009-1. Dynamic adaptive streaming over HTTP (DASH)-Part 1: Media presentation description and segment formats. Draft International Standard," 2011.

For citation:

T. V. Rykova, Towards the analysis of the performance measures of heterogeneous networks by means of two-phase queuing systems, Discrete and Continuous Models and Applied Computational Science 29 (3) (2021) 242–250. DOI: 10.22363/2658-4670-2021-29-3-242-250.

Information about the authors:

Rykova, Tatiana V. — Master of Science in Applied Mathematics and Informatics (PFUR), Master of Science in Information Technology (Tampere University of Technology), researcher at Fraunhofer Heinrich Hertz Institute (Berlin, Germany) (e-mail: tatiana.rykova@hhi.fraunhofer.de, ORCID: https://orcid.org/0000-0002-8561-7514)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2021-29-3-242-250

К анализу показателей эффективности гетерогенных сетей с помощью двухфазных систем массового обслуживания

Т. В. Рыкова

Российский университет дружбы народов
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия

Благодаря многоступенчатому характеру процессов передачи в гетерогенных мобильных сетях 4G, 5G, многофазные системы массового обслуживания становятся одним из наиболее подходящих способов анализа алгоритмов распределения ресурсов и исследования сетей. В этой статье приводится обзор нескольких научных работ, посвящённых моделированию гетерогенных сетей с помощью многофазных систем массового обслуживания, и упоминаются трудности, возникающие при этом типе аналитического анализа. Более того, несколько ранее исследованных моделей кратко представлены в качестве примера двухфазных систем конечной ёмкости и специальной структуры в дискретном времени, которые можно использовать для анализа схем распределения ресурсов на базе основных показателей производительности, полученных для беспроводных гетерогенных сетей. Одна из моделей представлена двухфазной тандемной очередью с групповым потоком поступающих запросов, а вторая — фазой сложной структуры, состоящей из параллельных конечных очередей. Вторая модель представляет собой двухфазную тандемную очередь с марковскими модулированными геометрическими процессами поступления и обслуживания на первом этапе и полным процессом обслуживания на втором этапе, что решает проблему межуровневой адаптации в гетерогенной сети.

Ключевые слова: двухфазная модель, система массового обслуживания, цепь Маркова, распределение ресурсов, гетерогенная сеть

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2021, 29 (3) 251–259
http://journals.rudn.ru/miph

Research article

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2021-29-3-251-259

Asymptotically accurate error estimates of exponential convergence for the trapezoidal rule

Aleksandr A. Belov[1, 2], **Valentin S. Khokhlachev**[1]

[1] *M. V. Lomonosov Moscow State University*
1, bld. 2, Leninskie Gory, Moscow, 119991, Russian Federation
[2] *Peoples' Friendship University of Russia (RUDN University)*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation

(received: August 19, 2021; accepted: September 9, 2021)

In many applied problems, efficient calculation of quadratures with high accuracy is required. The examples are: calculation of special functions of mathematical physics, calculation of Fourier coefficients of a given function, Fourier and Laplace transformations, numerical solution of integral equations, solution of boundary value problems for partial differential equations in integral form, etc. For grid calculation of quadratures, the trapezoidal, the mean and the Simpson methods are usually used. Commonly, the error of these methods depends quadratically on the grid step, and a large number of steps are required to obtain good accuracy. However, there are some cases when the error of the trapezoidal method depends on the step value not quadratically, but exponentially. Such cases are integral of a periodic function over the full period and the integral over the entire real axis of a function that decreases rapidly enough at infinity. If the integrand has poles of the first order on the complex plane, then the Trefethen–Weidemann majorant accuracy estimates are valid for such quadratures.

In the present paper, new error estimates of exponentially converging quadratures from periodic functions over the full period are constructed. The integrand function can have an arbitrary number of poles of an integer order on the complex plane. If the grid is sufficiently detailed, i.e., it resolves the profile of the integrand function, then the proposed estimates are not majorant, but asymptotically sharp. Extrapolating, i.e., excluding this error from the numerical quadrature, it is possible to calculate the integrals of these classes with the accuracy of rounding errors already on extremely coarse grids containing only ~ 10 steps.

Key words and phrases: trapezoidal rule, exponential convergence, error estimate, asymptotically sharp estimates

1. Introduction

Applied tasks. In many physical problems it is needed to calculate integrals, that cannot be obtained in terms of elementary functions. Here are some examples:

1) Calculation of special functions of mathematical physics: the Fermi–Dirac functions, which are equal to the moments of the Fermi distribution, the Gamma function, cylindrical functions and a number of others.
2) Calculation of the Fourier coefficients of a given function, Fourier and Laplace transform.
3) Numerical solution of integral equations, both well-posed and ill-posed.
4) Solving boundary value problems for partial differential equations (including eigenvalue problems) written in integral form, etc.

Calculation of quadratures. Trapezoidal rule, rectangle rule and Simpson's rule are commonly used for grid computation of quadratures. Usually the error of these methods quadratically depends on the grid step, and a large number of steps is needed to obtain good accuracy.

However, there are a number of cases when the error of the trapezoidal rule depends on the grid step exponentially, i.e. when the step is reduced by half, the number of correct signs of the numerical result is approximately doubled. This rate of convergence is similar to that of Newton's method. Two such cases are known. These are: the integral of the periodic function over the full period and the improper integral of a function that decreases rapidly enough at infinity.

If the integrand has first order poles on the complex plane, then for such quadratures there are majorant error estimates of Trefethen and Weidemann [1], see also [2]–[10]. In [11], [12] the generalization of Trefethen and Weidemann estimates is built for the case when the nearest pole of an integrand function is multiple.

In this paper, new error estimates of exponentially convergent quadratures of periodic functions over the full period are described. Integrand function can have an arbitrary number of poles of an integer order on the complex plane. If the mesh is detailed enough and the profile of the integrand resolved well, then the proposed estimates are not majorant, but asymptotically accurate.

It is possible to calculate the integrals of the indicated classes with the accuracy of round-off errors even on extremely coarse grids containing only ~ 10 steps by extrapolation (i.e., subtraction) of this error from the numerical value of the quadrature.

2. Exponentially convergent quadratures

One of the classes of exponentially convergent quadratures are integrals of periodic functions over the full period. By replacing $z = \exp(2\pi ix/X)$ we move from the integral over the period $[0, X]$ to the integral over the unit circle $|z| = 1$ on the complex plane. We choose the bypass direction of this circle counter clockwise. In [1], the following statement is formulated and proved:

Theorem 1. *Let $u(z)$ be analytic in the ring $R^{-1} < |z| < R$, where $R > 1$, and $|u(z)| < M_0$. We introduce a uniform grid on the unit circle $z_n = \exp(2\pi in/N)$, $n = \overline{0, N}$. Consider the integral and the trapezoidal rule quadrature*

$$I = \oint\limits_{|z|=1} u\left(z\right) \frac{dz}{iz}, \quad I_N = \frac{2\pi}{N} \sum_{n=1}^{N} u\left(z_n\right).$$

Then the estimate for the quadrature error holds

$$\delta = \left|I - I_N\right| \leqslant \frac{4\pi M_0}{R^N - 1}. \tag{1}$$

It is obvious that, by replacing $z = \exp\left(ix\right)$, theorem 1 holds for integral over full period of the function $u\left(\exp\left(ix\right)\right)$ on the real axis.

In the works [11], [12], it was shown that the dependence of the estimate (1) from N can be not majorant, but asymptotically accurate. This holds, if $u\left(z\right)$ has only first order pole type singularities, and R is taken such that the closest singularity to the unit circle lies on the boundary of the ring $R^{-1} < |z| < R$. In this case, the integrand function increases significantly if one approaches the singularity from inside the ring. Thereby, the constant M_0 loses its usual meaning from theorem 1. We carefully studied proof of the theorem 1 given in [1] and we found the possibility of significant strengthening the results of this theorem, under some additional conditions on the integrand function.

3. Calculating the error

Let us consider in detail the contour integral over the unit circle of a function that has one simple pole inside it and another simple one pole outside it. This case corresponds to the integral considered in [1]. Suppose the point a_1 is inside, and the point a_2 is outside $|z| = 1$ and, the function $u\left(z\right)$ is analytic in the ring $R^{-1} < |z| < R$, where $R = \min\left\{1/\left|a_1\right|, \left|a_2\right|\right\}$. Then the integral has the form

$$G = \oint\limits_{|z|=1} g\left(z\right) dz = \oint\limits_{|z|=1} \frac{u\left(z\right)}{\left(z - a_1\right)\left(z - a_2\right)} dz = 2\pi i \frac{u\left(a_1\right)}{\left(a_1 - a_2\right)}.$$

We make one assumption for the sake of simplifying the calculations. Its effect on the result is weak. Let $u\left(z\right) = 1$, then we rewrite the integrand function in this form

$$g\left(z\right) = \frac{1}{\left(z - a_1\right)\left(z - a_2\right)} = \frac{1}{\left(a_1 - a_2\right)\left(z - a_1\right)} + \frac{1}{\left(a_2 - a_1\right)\left(z - a_2\right)}.$$

Now we decompose each fraction in the Laurent series as the sum of the geometric progression

$$g\left(z\right) = \frac{1}{\left(a_1 - a_2\right)} \sum_{k_1=0}^{\infty} \frac{a_1^{k_1}}{z^{k_1+1}} - \frac{1}{\left(a_2 - a_1\right)} \sum_{k_2=0}^{\infty} \frac{z^{k_2}}{a_2^{k_2+1}}.$$

We use the grid z_n, $n = \overline{0, N}$, which is introduced in theorem 1. Our goal is to obtain an explicit expression for the grid step $\Delta z_n = z_{n+1} - z_n$

$$\Delta z_n = \exp\left(\frac{2\pi i\,(n+1)}{N}\right) - \exp\left(\frac{2\pi i n}{N}\right) \underset{N\to\infty}{=} z_n\left(\frac{2\pi i}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)\right). \quad (2)$$

Discarding the $\mathcal{O}\left(N^{-2}\right)$ term in the expression for the grid step, we write the trapezoidal rule quadrature in the following form

$$G_N = \sum_{n=0}^{N-1} g\left(z_n\right)\Delta z_n = \frac{2\pi i}{N}\sum_{n=0}^{N-1} g\left(z_n\right) z_n.$$

We substitute the representation of $g\left(z_n\right)$ by the sum of the series in the quadrature and then swap the series and the finite sum. Last step is allowed due to absolute convergence of the resulting double number series (each member of the double series of modules can be estimated by the corresponding member of an infinitely decreasing geometric progression, which has finite sum). The following expression for the quadrature formula is obtained

$$G_N = \frac{2\pi i}{N(a_1 - a_2)}\left[\sum_{s_1=0}^{\infty}\sum_{n=0}^{N-1}\frac{a_1^{s_1}}{z_n^{s_1}} + \sum_{s_2=0}^{\infty}\sum_{n=0}^{N-1}\frac{z_n^{s_2+1}}{a_2^{s_2+1}}\right]. \quad (3)$$

To perform these transformations, we need the following well known result

$$\sum_{n=0}^{N-1}\exp\left(\pm 2\pi i\frac{nk}{N}\right) = \begin{cases} N, & k \text{ is a multiple of } N,\\ 0, & \text{otherwise.}\end{cases}$$

We convert the second sum in square brackets in the formula (3)

$$\sum_{s_2=0}^{\infty}\sum_{n=0}^{N-1}\frac{z_n^{s_2+1}}{a_2^{s_2+1}} = \left\{\begin{matrix} (s_2+1) \text{ is a multiple of } N,\\ (s_2+1) = Np_2,\\ p_2 = \overline{1,\infty}\end{matrix}\right\} =$$

$$= N\sum_{p_2=1}^{\infty}\frac{1}{a_2^{Np_2}} = N\frac{1/a_2^N}{1 - 1/a_2^N}.$$

We convert the first sum in (3)

$$\sum_{s_1=0}^{\infty}\sum_{n=0}^{N-1}\frac{a_1^{s_1}}{z_n^{s_1}} = \left\{\begin{matrix} s_1 \text{ is a multiple of } N,\\ s_1 = Np_1, \quad p_1 = \overline{0,\infty}\end{matrix}\right\} = N\sum_{p_1=0}^{\infty} a_1^{Np_1} = N\frac{1}{1 - a_1^N}.$$

We get

$$G_N = \frac{2\pi i}{(a_1 - a_2)}\left[\frac{1}{1 - a_1^N} + \frac{1}{a_2^N - 1}\right].$$

Finally, we calculate the quadrature error

$$\Delta_N = G - G_N = 2\pi i \left(\frac{1}{(a_1 - a_2)} \left[1 - \frac{1}{1 - a_1^N} \right] + \frac{1}{(a_2 - a_1)} \left[\frac{1}{a_2^N - 1} \right] \right).$$

The obtained result can be easily generalized to the case when the function $u(z)$ does not equal identically to one. The derivation is similar but far too cumbersome. Let us formulate the final result.

Theorem 2. *Let the point $z = a_1$ be inside the unit circle, and let the point $z = a_2$ be outside of it. Let the function $u(z)$ be analytic on the entire complex plane, with the possible exception of an infinitely distant point, and $u \neq 0$ at the points $z = a_{1,2}$. Then the trapezoidal rule for the integral G has the following error estimation*

$$\Delta_N = G - G_N = \\ = 2\pi i \left(\frac{u(a_1)}{(a_1 - a_2)} \left[1 - \frac{1}{1 - a_1^N} \right] + \frac{u(a_2)}{(a_2 - a_1)} \left[\frac{1}{a_2^N - 1} \right] \right). \quad (4)$$

Estimate (4) is not majorant, but asymptotically accurate. The only one approximation that was made is contained in the approximate expression for the grid step (2).

4. Validation

Calculations were carried out with the test integral having a known value

$$J = \oint\limits_{|z|=1} \frac{\sin(z)}{(z - a_1)(z - a_2)} dz = 2\pi i \frac{\sin(a_1)}{(a_1 - a_2)}, \quad (5)$$

where $a_1 = 0.6 + 0.6i$ and $a_2 = 2 - i$. In this case, $1/|a_1| \approx 1.2$ and $|a_2| \approx 2.2$, so the value R from theorem 1 equals $1/|a_1|$. During the calculations, the following information was obtained: actual error, the Trefethen–Weidemann estimate (1), our estimate (4) and the error after extrapolation.

The figure 1 shows quadrature error versus number of grid steps in the semi-logarithmic scale. Here, the black dots represent the actual error, the white circles represent our estimate, and the black squares represent the Trefethen–Weidemann estimate with the constant $M_0 = 1$. Recall that this constant loses its meaning from theorem 1, if the singularity lies on the boundary of the ring.

The plot shows that our estimate coincides with the actual error already at $N > 4$. The Trefethen–Weidemann estimate does not represent the initial part of the curve. It describes the curve starting from $N \cong 15$. This estimate is asymptotically accurate in N, but the true value of the constant M_0 is unknown. Therefore, the Trefethen–Weidemann estimate cannot be used for extrapolation. Thereby, the error estimate constructed in this paper is much stronger than the Trefethen–Weidemann estimate.

These conclusions are also confirmed by the figure 2. Here, we plot the ratio of error estimates to actual accuracy versus number of grid steps. The number 1 corresponds to the Trefethen–Weidemann estimate and the number 2 is for our estimate. It can be seen that when $N > 4$ our estimate is almost indistinguishable from the actual error. Therefore, it can be excluded from the quadrature (i.e. extrapolated). This dramatically increases the accuracy of the calculation. One can also see that the Trefethen–Weidemann estimate significantly less accurate in assessing the dependence of the error on the number of nodes: the corresponding relation goes out to a constant on the much more detailed grids than the estimate (4).

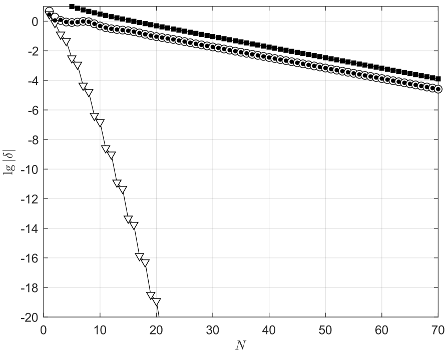


Figure 1. Graph of convergence of the trapezoidal formula. Symbols are described in the text

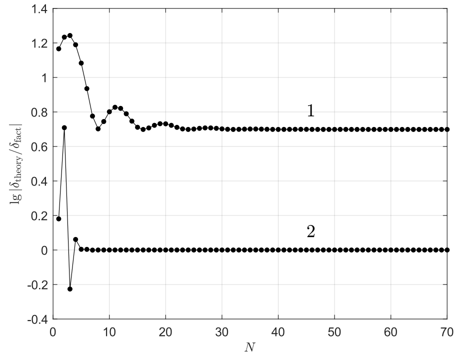


Figure 2. Ratio of theoretical estimates (1) and (4) for the integral (5) to actual accuracy versus number of grid steps. Symbols see in the text

5. Extrapolation of the error

Let us exclude the error (4) from the calculated quadrature by the formula

$$\widetilde{G}_N = G_N + \Delta_N. \tag{6}$$

This is equivalent to introducing some new quadrature formula. The error is shown in the figure 1 by white triangles. One can see that the speed of convergence of the quadrature (6) radically exceeds even the exponential one. The accuracy of round-off errors is achieved already at $N \sim 15$, which is ~ 10 times less than for the trapezoidal rule. Labor intensity of such computation is comparable to the complexity of explicit formulas. This approach is essentially new and exceeds the world level.

6. Conclusion

The described method is a powerful tool for solving physical problems. If one can find transformation of variables that reduce integral to one of the considered types, then the calculations are accelerated thousands of times. In this paper 1) a fundamentally new estimate of the error of quadrature is constructed, it is asymptotically accurate. 2) Extrapolation procedure is proposed, which provides calculation of the quadrature with the accuracy of unit errors rounding, and it is already performed on very rough grids with the number of steps from 5–15.

Acknowledgments

This work was supported by grant MK-3630.2021.1.1.

References

[1] L. N. Trefethen and J. A. C. Weideman, "The exponentially convergent trapezoidal rule," *SIAM Review*, vol. 56, no. 3, pp. 385–458, 2014. DOI: 10.1137/130932132.

[2] J. Mohsin and L. N. Trefethen, "A trapezoidal rule error bound unifying the Euler–Maclaurin formula and geometric convergence for periodic functions," in *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 470, 2014, p. 20 130 571. DOI: 10.1098/rspa.2013.0571.

[3] J. A. C. Weideman, "Numerical integration of periodic functions: A few examples," *The American Mathematical Monthly*, vol. 109, no. 1, pp. 21–36, 2002. DOI: 10.2307/2695765.

[4] N. Eggert and J. Lund, "The trapezoidal rule for analytic functions of rapid decrease," *Journal of Computational and Applied Mathematics*, vol. 27, no. 3, pp. 389–406, 1989. DOI: 10.1016/0377-0427(89)90024-1.

[5] H. Al Kafri, D. J. Jeffrey, and R. M. Corless, "Rapidly convergent integrals and function evaluation," *Lecture Notes in Computer Science*, vol. 10693, pp. 270–274, 2017. DOI: 10.1007/978-3-319-72453-9_20.

[6] J. Waldvogel, "Towards a general error theory of the trapezoidal rule," in *Springer Optimization and Its Applications*. 2010, vol. 42, pp. 267–282. DOI: 10.1007/978-1-4419-6594-3_17.

[7] E. T. Goodwin, "The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(x)e^{-x^2}dx$," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 45, no. 2, pp. 241–245, 1949. DOI: 10.1017/S0305004100024786.

[8] N. N. Kalitkin and S. A. Kolganov, "Quadrature formulas with exponential convergence and calculation of the Fermi–Dirac integrals," *Doklady Mathematics*, vol. 95, no. 2, pp. 157–160, 2017. DOI: 10.1134/S1064562417020156.

[9] N. N. Kalitkin and S. A. Kolganov, "Refinements of precision approximations of Fermi–Dirak functions of integer indices," *Mathematical Models and Computer Simulations*, vol. 9, no. 5, pp. 554–560, 2017. DOI: 10.1134/S2070048217050052.

[10] N. N. Kalitkin and S. A. Kolganov, “Computing the Fermi–Dirac functions by exponentially convergent quadratures,” *Mathematical Models and Computer Simulations*, vol. 10, no. 4, pp. 472–482, 2018. DOI: 10.1134/S2070048218040063.

[11] A. A. Belov, N. N. Kalitkin, and V. S. Khokhlachev, “Improved error estimates for an exponentially convergent quadratures [Uluchshennyye otsenki pogreshnosti dlya eksponentsial’no skhodyashchikhsya kvadratur],” *Preprints of IPM im. M. V. Keldysh*, no. 75, 2020, in Russian. DOI: 10.20948/prepr-2020-75.

[12] V. S. Khokhlachev, A. A. Belov, and N. N. Kalitkin, “Improvement of error estimates for exponentially convergent quadratures [Uluchsheniye otsenok pogreshnosti dlya eksponentsial’no skhodyashchikhsya kvadratur],” *Izv. RAN. Ser. fiz.*, vol. 85, no. 2, pp. 282–288, 2021, in Russian. DOI: 10.31857/S0367676521010166.

For citation:
A. A. Belov, V. S. Khokhlachev, Asymptotically accurate error estimates of exponential convergence for the trapezoidal rule, Discrete and Continuous Models and Applied Computational Science 29 (3) (2021) 251–259. DOI: 10.22363/2658-4670-2021-29-3-251-259.

Information about the authors:

Belov, Aleksandr A. — Candidate of Physical and Mathematical Sciences, Assistant professor of Department of Applied Probability and Informatics of Peoples’ Friendship University of Russia (RUDN University); Researcher of Faculty of Physics, M. V. Lomonosov Moscow State University (e-mail: aa.belov@physics.msu.ru, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0002-0918-9263, ResearcherID: Q-5064-2016, Scopus Author ID: 57191950560)

Khoklachev, Valentin S. — Master’s degree student of Faculty of Physics, M. V. Lomonosov Moscow State University (e-mail: valentin.mycroft@yandex.ru, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0002-6590-5914)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2021-29-3-251-259

Асимптотически точные оценки экспоненциальной сходимости для формулы трапеций

А. А. Белов[1, 2], **В. С. Хохлачев**[1]

[1] *Московский государственный университет им. М. В. Ломоносова*
Ленинские горы, д. 1, стр. 2, Москва, 119991, Россия
[2] *Российский университет дружбы народов*
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия

Во многих прикладных задачах требуется экономичное вычисление квадратур с высокой точностью. Примерами являются: вычисление специальных функций математической физики, расчёт коэффициентов Фурье заданной функции, преобразования Фурье и Лапласа, численное решение интегральных уравнений, решение краевых задач для уравнений в частных производных в интегральной форме и т.д. Для сеточного вычисления квадратур обычно используют методы трапеций, средних и Симпсона. Обычно погрешность этих методов зависит от шага степенным образом, и для получения хорошей точности требуется большое число шагов. Однако существует ряд случаев, когда погрешность метода трапеций зависит от величины шага не квадратично, а экспоненциально. Такими случаями являются интеграл от периодической функции по полному периоду и интеграл по всей числовой прямой от функции, достаточно быстро убывающей на бесконечности. Если подынтегральная функция имеет полюса первого порядка в комплексной плоскости, то для таких квадратур справедливы мажорантные оценки точности Трефетена и Вайдемана.

В работе построены новые оценки погрешности экспоненциально сходящихся квадратур от периодических функций по полному периоду. Подынтегральная функция может иметь произвольное число полюсов целого порядка на комплексной плоскости. Если сетка достаточно подробная (разрешает профиль подынтегральной функции), то предлагаемые оценки являются не мажорантными, а асимптотически точными. Экстраполируя, то есть исключая эту погрешность из численной квадратуры, можно вычислять интегралы указанных классов с точностью ошибок округления уже на чрезвычайно грубых сетках, содержащих всего ~ 10 шагов.

Ключевые слова: формула трапеций, экспоненциальная сходимость, оценки точности, асимптотически точные оценки

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2021, 29 (3) 260–270
http://journals.rudn.ru/miph

Research article

UDC 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2021-29-3-260-270

Shifted Sobol points and multigrid Monte Carlo simulation

Aleksandr A. Belov[1, 2], **Maxim A. Tintul**[1]

[1] *M. V. Lomonosov Moscow State University*
1, bld. 2, Leninskie Gory, Moscow, 119991, Russian Federation
[2] *Peoples' Friendship University of Russia (RUDN University)*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation

(received: August 19, 2021; accepted: September 9, 2021)

Multidimensional integrals arise in many problems of physics. For example, moments of the distribution function in the problems of transport of various particles (photons, neutrons, etc.) are 6-dimensional integrals. When calculating the coefficients of electrical conductivity and thermal conductivity, scattering integrals arise, the dimension of which is equal to 12. There are also problems with a significantly large number of variables. The Monte Carlo method is the most effective method for calculating integrals of such a high multiplicity. However, the efficiency of this method strongly depends on the choice of a sequence that simulates a set of random numbers. A large number of pseudo-random number generators are described in the literature. Their quality is checked using a battery of formal tests. However, the simplest visual analysis shows that passing such tests does not guarantee good uniformity of these sequences. The magic Sobol points are the most effective for calculating multidimensional integrals. In this paper, an improvement of these sequences is proposed: the shifted magic Sobol points that provide better uniformity of points distribution in a multidimensional cube. This significantly increases the cubature accuracy. A significant difficulty of the Monte Carlo method is a posteriori confirmation of the actual accuracy. In this paper, we propose a multigrid algorithm that allows one to find the grid value of the integral simultaneously with a statistically reliable accuracy estimate. Previously, such estimates were unknown. Calculations of representative test integrals with a high actual dimension up to 16 are carried out. The multidimensional Weierstrass function, which has no derivative at any point, is chosen as the integrand function. These calculations convincingly show the advantages of the proposed methods.

Key words and phrases: multidimensional integral, Monte Carlo method, Sobol points, multigrid calculation, a posteriori error estimates

1. Introduction

Integrals of multivariate functions occur in many areas of physics. Here are some examples. The transfer of neutrons, photons and other particles in the medium is described by the equation for the distribution function; this function depends on three coordinates of the medium and three components of the particle velocity vector, that is, the number of variables is six. To determine the coefficients of thermal conductivity or electrical conductivity of a medium, it is necessary to calculate the collision integrals; they include components of the velocity vectors before the moment of collision and after the moment of collision. The total number of variables in such an integral is twelve. Problems also arise with a significantly larger number of variables.

In the simplest formulation, the calculation of the integral in the unit p-dimensional cube V is considered. $x = (x_1, x_2, \dots, x_p)$ is p-dimensional vector. Our aim is to calculate the following integral:

$$I \equiv \int\limits_V f(x)dx = \int\limits_0^1 \dots \int\limits_0^1 f(x_1, x_2, \dots, x_p)dx_1 dx_2 \dots dx_p.$$

The accuracy of numerical grid methods drops rapidly with the increase of dimension p. In order to obtain acceptable accuracy, more and more points have to be taken, which makes the calculations exorbitantly laborious and very time consuming. Due to this fact, the local Monte Carlo method is used for high dimensions ($p > 3$). It involves the use of random numbers, which are mathematical abstraction. In practice, however, one has to use sequences that only imitate random numbers. Performance of the method strongly depends on the choice of such a sequence.

Calculations of the representative test integrals show that to obtain good accuracy the most important is the uniformity of the points' distribution and not its randomness. The most effective are Sobol sequences with the so-called "magic" numbers of points $N = 2^n$, $n = 0, 1, \dots$.

In this work, the following results are obtained. Firstly, shifted Sobol points are proposed. It is a modification that improves uniformity of the point distribution and increases the accuracy of cubatures. Secondly, a multigrid strategy that gives a posteriori estimate of the accuracy is constructed. The advantages of the proposed algorithms are illustrated with representative test examples.

2. Pseudorandom points

For the local Monte Carlo method, N random points x_j are selected in the cube V; in this case, the number N can be arbitrary, in contrast to cubature formulae on regular grids. The cubature formula

$$I_N \equiv \frac{1}{N}\sum_{j=1}^{N} f(x_j) \tag{1}$$

is similar to the formula for mean Riemann sum. However, the estimate of its error ΔN turns out to be radically different

$$\Delta_N \equiv I - I_N \sim \sqrt{DfN^{-1/2}}, \quad Df = \int\limits_V f^2(x)dx - \left[\int\limits_V f(x)dx\right]^2. \tag{2}$$

Here Df is variance. The estimate of the error is not majorant, but probabilistic: magnitude of the error is distributed according to the Gaussian law with the standard specified in the formula. The error does not exceed the standard deviation with a probability of 0.68.

The error estimate (2) does not depend on the dimension p. Random points are inferior in accuracy to regular grids at $p = 1$ or $p = 2$. Already at $p = 4$, the dependence of the error on N for random points and regular grids is the same. With further increase in dimension, random points turns out to be more advantageous; advantage increases rapidly as dimension p grows.

Formulae (2) assume that random points x_j have uniform distribution density in the cube V and are not correlated. However, no rigorous mathematical methods for constructing such points have been found. A number of mathematical algorithms have been proposed; the resulting points are called pseudorandom. An extensive literature is devoted to the construction of pseudorandom points, for example, [1]–[13]. The following generators are most common in the literature:

— Mersenne twister and SIMD-oriented fast Mersenne twister;
— Multiplicative congruential generator;
— 64-bit multiplicative lagged Fibonacci generator;
— combined multiple recursive generator;
— generator Philo4x32;
— generator Threefry4x64;
— Marsaglia's SHR3 shift-register generator;
— modified Subtract-with-Borrow generator;
— modified Lehmer sequence.

These generators are implemented in many commercial packages (for example, Matlab).

The quality of each sequence of pseudorandom numbers is checked using some sets of tests based on the theory of probability [14]–[17]. But no set of tests can be complete and comprehensive. Therefore, such checks are limited. Even the simplest visual tests show that widespread sequences do not provide a sufficiently good uniformity of filling the unit square [18], [19]. The question of the influence of such unevenness on the actual error of cubatures remains insufficiently clear.

3. Sobol points

To construct the Sobol sequence, a set of the so-called direction numbers should be selected. There is some ambiguity in the selection of initial direction numbers. In early works [1], direction number tables were constructed for dimensions $p \leqslant 13$ and numbers $n \leqslant 20$ (total number of points $N \leqslant 2^{20}$). Later, direction numbers for higher p and N were constructed [20]. However, the direction numbers were also changed. The program is currently available at **21**. The open access option contains $p \leqslant 50$ and $n \leqslant 31$ $(N \approx 2 \cdot 10^9)$. The commercial version of the program has $p \leqslant 2^{16} - 1$.

It is important to note that the Sobol sequences are constructed separately for each p. It is impossible to obtain a sequence of fewer dimensions from p-dimensional Sobol sequence. This also applies to magic segments of the Sobol sequences.

The Sobol cubature formula has the same form as (2). But the estimate of its error is not entirely clear. The distribution of points only for magic N approaches uniform in properties. For intermediate N, it is obtained by discarding some of the points and loses the property of uniformity. Therefore, only magic N should be used for cubatures.

Various attempts have been made to generalize the Sobol sequences. However, the search for optimal variants of such generalizations invariably led again to the Sobol sequences. Therefore, such generalizations need to be treated with caution.

4. Shifted Sobol points

The arrangement of the Sobol points is somewhat asymmetrical. For example, if number of points $N = 2^n$ is taken, then the arithmetic mean of all points projections on any axis will not be 0.5, but $0.5\,(1 - 1/N)$. Obviously, this asymmetry is not favourable for obtaining good cubature accuracy.

In the figure 1, black circles show two-dimensional Sobol points for the first magic numbers. For $n = 0$, the only point lies in the corner of the unit square. Calculation of the cubature over this point gives a formula of the first order accuracy. However, if this point is shifted by 0.5 along each coordinate, then the cubature over the shifted point (light circle) has the second order of accuracy. For the case $n = 1$, two points are located one in the corner of the square and one in the center, which will also give the first order of accuracy. But if these two points are shifted by 0.25 along each coordinate, then the cubature error obviously decreases. Therefore, a general shift principle for any number of dimensions can be proposed:

If $N = 2^n$***, then add to all coordinates of all points*** $(2N)^{-1}$***.***

It is advisable to apply this shift only for magic Sobol numbers. In this case, the shifts are different for different N.

5. Multigrid calculation

Test calculations show that the actual error decreases as $O\left(N^{-1}\right)$. This suggests that it is possible to approximate the integral (and hence its error) as a function of N. However, this approximation cannot be smooth, such as Richardson's interpolation approximation for grid methods. In this case, the points are obtained by statistical methods, therefore, their processing must be carried out using the root-mean-square approximation. To do this, the type of approximation must be chosen and some weights to the points need to be assigned.

As a working hypothesis, the law of decreasing error $\Delta_N \sim N^{-1}$ was assumed. But since the nature of the error becomes clearly statistical with increasing p, the standard deviation of these errors was assumed to be proportional to $N^{-1/2}$. This is the weight used for approximation.

The following multigrid procedure is proposed. The calculation with magic $N = 2^n$, $n = 10, 11, ...$ is performed. As a result, a sequence of values of the integral $\{I_N\}$ is obtained. Now this sequence can be approximated by the method of least squares

$$I_N \approx a + bN^{-1}. \qquad (3)$$

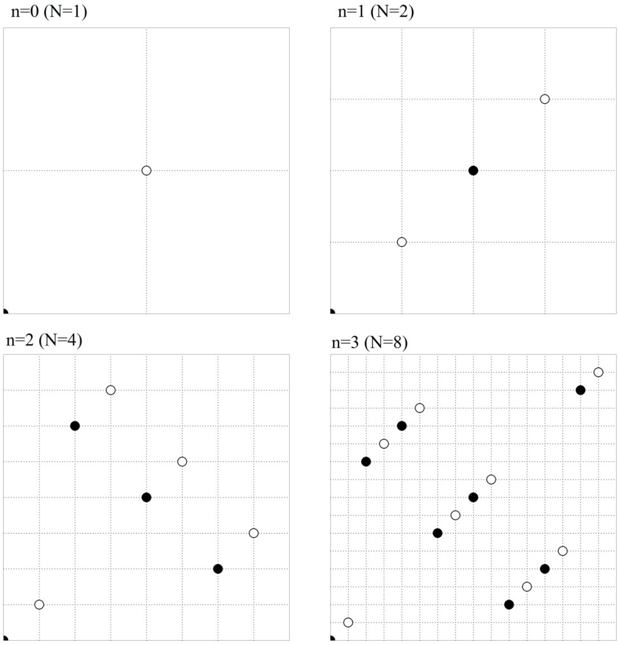


Figure 1. Sobol magic points for $p = 2$: points – unbiased, circles – shifted; the n values are indicated near the squares

Here a is the refined value of the integral. At the same time, the standard deviation σ_a for the value a is calculated. This standard deviation is a statistical estimate of the accuracy for the found value of the integral.

Note that the beginning of the sequence $\{I_N\}$ corresponding to $n = 0, 1, \dots, 9$ is not taken into account in approximation (3), since these grids are not detailed enough, and the rate of decrease of the error does not yet correspond to $O\left(N^{-1}\right)$.

6. Test integral

It is expedient to carry out numerical experiments on multidimensional integrals over the unit cube, the exact values of which are known. Then the error of the numerical calculation can be directly determined and its behaviour can be studied. Further, requirements that are appropriate for the integrand are discussed.

In multidimensional problems, the concept of the effective dimension of a function is used. For example, consider two functions:

$$f(x) = \prod_{j=1}^{p} f_j\left(x_j\right) \tag{4}$$

and

$$f(x) = f_1\left(\sum_{j=1}^{p} \alpha_j x_j\right),$$

where all $f_j(x_j)$ are essentially different from constants. In the first function, all variables are equally important, and the effective dimension of the function is p. The second function depends on only one combination of variables, so its effective dimension is 1. The higher the effective dimension of the function, the more difficult the problem. Therefore, the most difficult functions are of the first type.

Suppose that for a product function each f_j differs substantially from zero only on a segment of length β of its unit edge. Then the product of one-dimensional functions will differ significantly from zero in the volume β^p. If β is small, then as p increases, the volume β^p decreases rapidly; for example, for $\beta = 0.1$ and $p = 10$ the value $\beta^p = 10^{-10}$. In this case, to obtain acceptable accuracy, any Monte Carlo method will require the number of nodes $N \gg \beta^{-p}$. It can be seen that in order for the number of points to be reasonable, β should be taken close to one.

Taking these considerations into account, a test of the form (4) have been chosen. It is not easy, despite its seeming simplicity. All f_j are assumed to be the same and equal to the Weierstrass functions

$$f_j\left(x_j\right) = \sum_{n=0}^{\infty} b^n \cos\left(a^n \pi x_j\right), \tag{5}$$

where a is an arbitrary odd number that is not equal to one, and b is a positive number less than one. It is known that under the conditions $ab \geqslant 1$, $a > 1$, the Weierstrass function is continuous, but has no derivative at any point. This test is extremely difficult. The Weierstrass function is shown in the figure 2.

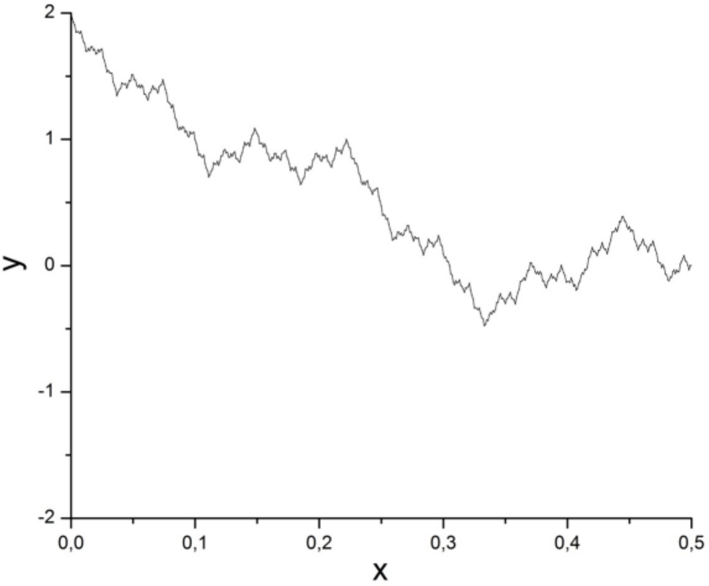


Figure 2. Weierstrass function with $a = 3, b = 0.5$

Taking into account the symmetry of the Weierstrass function, the integration is carried out over a cube with sides $x_j \in [0, 0.5]$. For convenience, the Weierstrass function is normalized, the normalization condition is

$$\int\limits_V f(x)dx = 1. \tag{6}$$

7. Calculation results

The integral of the multidimensional Weierstrass function (4), (5) was calculated using three qualitatively different approaches: regular cubature on trapezoidal formulae, the classical Monte Carlo method using the Mersenne twister and shifted Sobol magic points.

These three approaches are compared in terms of the error magnitude with a fairly modest number of points $N = 2^{20}$. The logarithms of the errors depending on the dimension are shown in the figure 3. Let us analyze the curves.

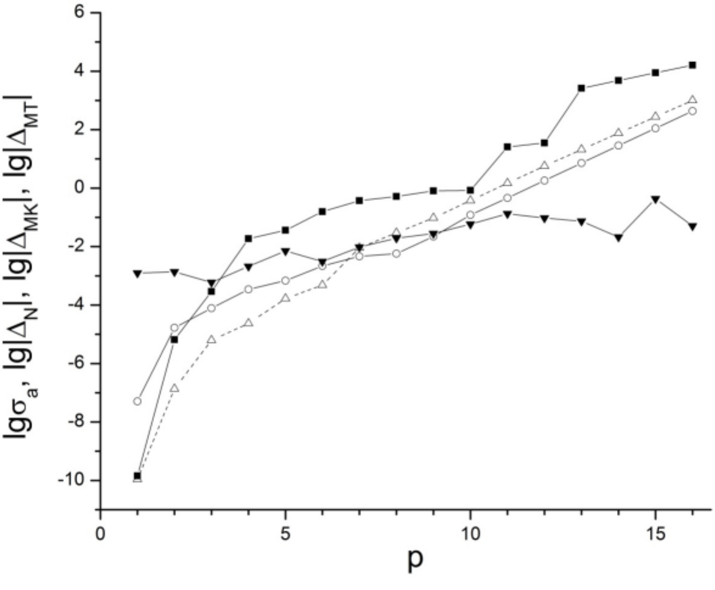


Figure 3. Logarithm of the relative error in calculating the integral of the Weierstrass function for $N = 2^{20}$: light triangle is Δ_N, circle is σ_a for the shifted Sobol points, black inverted triangle corresponds to Mersenne twister, black square is for trapezoidal method

Mersenne twister. Beginning with dimension $p = 11$, the curve corresponding to the Mersenne twister lies below all. Despite the good accuracy, there are no means to confirm it. An attempt to apply the root-mean-square approximation (3) to the Mersenne twister was unsuccessful: the values of σ_a turn out to be either larger or smaller than the actual error depending on the dimension p, and the difference can be significant.

The standard deviation $(Df/N)^{1/2}$ can serve as an error estimate of the Mersenne twister, but the calculation of the variance for some integrals can be problematic. In addition, the performance of the Monte Carlo method is highly dependent on the choice of a sequence that simulates random numbers, so the standard and actual error can vary greatly.

In general, the value of $\lg|\Delta_{MK}|$ lies in the range from -3.5 to 0 and slowly increases with increasing dimension p.

Trapezoidal formula. Its error is determined by the formula

$$|\Delta_N| \leqslant \frac{1}{12k^2}\max\left|\frac{d^2 f_j}{dx^2}\right|, \tag{7}$$

where $k = N^{1/p}$ is the number of nodes along each coordinate. Thus, its error is $O\left(N^{-2/p}\right)$; accuracy should decrease rapidly with increasing . The corresponding curve (black square marker in Fig. 3) illustrates good accuracy $\lg|\Delta_{MT}| \approx -5.2$ at $p = 2$; this is much more accurate than the classical Monte Carlo method. However, with increasing p, the error rapidly increases, and already at $p \geqslant 4$ it exceeds the error of the Monte Carlo method. At even higher dimensions, the trapezoidal method quickly becomes uncompetitive.

Sobol sequence. Despite the fact that for high dimensions p the Mersenne twister shows the best result, the shifted Sobol points have a reasonable estimate of the accuracy. It is the standard deviation σ_a. Thus, even in complex problems, the actual accuracy can be estimated a posteriori using σ_a, the number of points can be increased and the calculation can be repeated. This is especially important for multidimensional integrals with an unknown exact answer.

8. Conclusion

The magic Sobol points are the most effective for calculating multidimensional integrals. In this paper, an improvement of these sequences is proposed. They are called the shifted Sobol magic points, which provide a more uniform distribution of points in a multidimensional cube. This significantly increases the accuracy of cubatures.

A significant difficulty with Monte Carlo methods is the a posteriori confirmation of the actual accuracy. In this paper, a multigrid algorithm is proposed that allows to find the grid value of the integral simultaneously with a statistically reliable estimate of its accuracy. Previously, such estimates were unknown.

Calculations of representative test integrals with high actual dimension p (up to $p = 16$) are carried out. Smooth integrands were considered, as well as the multidimensional Weierstrass function having no derivative at any point. These calculations convincingly show the advantages of the proposed methods.

Acknowledgments

This work was supported by grant MK-3630.2021.1.1.

References

[1] I. M. Sobol, *Numerical Monte-Carlo methods [Chislennyye metody Monte-Karlo]*. Moscow: Nauka, 1973, In Russian.

[2] D. E. Knuth, *The art of computer programming*, 3rd ed. Reading, Massachusetts: Addison-Wesley, 1997, vol. 2.

[3] G. S. Fishman, *Monte Carlo: concepts, algorithms and applications.* Berlin: Springer, 1996. DOI: 10.1007/978-1-4757-2553-7.

[4] M. Matsumoto and T. Nishimura, "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 8, no. 1, pp. 3–30, 1998. DOI: 10.1145/272991.272995.

[5] T. Nishimura, "Tables of 64-bit Mersenne twisters," *ACM Transactions on Modeling and Computer Simulation*, vol. 10, no. 4, pp. 348–357, 2000. DOI: 10.1145/369534.369540.

[6] "Mersenne Twister Home Page." (2021), [Online]. Available: http://www.math.sci.hiroshima-u.ac.jp/m-mat/MT/emt.html.

[7] S. K. Park and K. W. Miller, "Random number generators: good ones are hard to find," *Communications of the ACM*, vol. 31, no. 10, pp. 1192–1201, 1998. DOI: 10.1145/63039.63042.

[8] M. Mascagni and A. Srinivasan, "Parameterizing parallel multiplicative Lagged–Fibonacci generators," *Parallel Computing*, vol. 30, pp. 899–916, 2004. DOI: 10.1016/j.parco.2004.06.001.

[9] P. L'Ecuyer, "Good parameter sets for combined multiple recursive random number generators," *Operations Research*, vol. 47, no. 1, pp. 159–164, 1999. DOI: 10.1287/opre.47.1.159.

[10] J. K. Salmon, M. A. Moraes, R. O. Dror, and D. E. Shaw, "Parallel random numbers: as easy as 1, 2, 3," *2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pp. 1–12, 2011. DOI: 10.1145/2063384.2063405.

[11] G. Marsaglia and W. W. Tsang, "The ziggurat method for generating random variables," *Journal of Statistical Software*, vol. 5, pp. 1–7, 2000. DOI: 10.18637/jss.v005.i08.

[12] G. Marsaglia and A. Zaman, "A new class of random number generators," *Annals of Applied Probability*, vol. 1, no. 3, pp. 462–480, 1991. DOI: 10.1214/aoap/1177005878.

[13] B. A. Wichmann and I. D. Hill, "An efficient and portable pseudo-random number generator," *Applied Statistics*, vol. 31, no. 2, pp. 188–190, 1982. DOI: 10.2307/2347988.

[14] E. A. Tsvetkov, "Empirical tests for statistical properties of some pseudorandom number generators," *Mathematical Models and Computer Simulations*, vol. 3, pp. 697–705, 2011. DOI: 10.1134/S20700-4821106010X.

[15] “The Marsaglia Random Number CDROM including the Diehard Battery of Tests of Randomness.” (2021), [Online]. Available: `http://ftpmirror.your.org/pub/misc/diehard/`.

[16] P. L’Ecuyer and R. Simard, “TestU01: A C library for empirical testing of random number generators,” *ACM Transactions on Mathematical Software (TOMS)*, vol. 33, no. 4, pp. 1–40, 2007. DOI: `10.1145/1268776.1268777`.

[17] L. E. Bassham, A. L. Rukhin, J. Soto, J. R. Nechvatal, M. E. Smid, S. D. Leigh, M. Levenson, M. Vangel, N. A. Heckert, and D. L. Banks, “A statistical test suite for random and pseudorandom number generators for cryptographic applications,” National Institute of Standards and Technology, NIST Special Publication, Gaithersburg, MD, Tech. Rep., 2010.

[18] A. A. Belov, N. N. Kalitkin, and M. A. Tintul, “Visual verification of pseudo-random number generators [Vizual’naya verifikatsiya generatorov psevdosluchaynykh chisel],” Keldysh IAM Preprints, Moscow, Tech. Rep. 137, 2019, In Russian. DOI: `10.20948/prepr-2019-137`.

[19] A. A. Belov, N. N. Kalitkin, and M. A. Tintul, “Unreliability of pseudorandom number generators,” *Computational Mathematics and Mathematical Physics*, vol. 60, no. 11, pp. 1747–1753, 2020. DOI: `10.1134/S0965542520110044`.

[20] I. M. Sobol, “Uniformly distributed sequences with additional uniformity properties,” *USSR Computational Mathematics and Mathematical Physics*, vol. 16, no. 5, pp. 236–242, 1976. DOI: `10.1016/0041-5553(76)90154-3`.

For citation:

A. A. Belov, M. A. Tintul, Shifted Sobol points and multigrid Monte Carlo simulation, Discrete and Continuous Models and Applied Computational Science 29 (3) (2021) 260–270. DOI: 10.22363/2658-4670-2021-29-3-260-270.

Information about the authors:

Belov, Aleksandr A. — Candidate of Physical and Mathematical Sciences, Assistant professor of Department of Applied Probability and Informatics of Peoples’ Friendship University of Russia (RUDN University); Researcher of Faculty of Physics, M. V. Lomonosov Moscow State University (e-mail: `aa.belov@physics.msu.ru`, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0002-0918-9263, ResearcherID: Q-5064-2016, Scopus Author ID: 57191950560)

Tintul, Maxim A. — Master’s degree student of Faculty of Physics, M. V. Lomonosov Moscow State University (e-mail: `maksim.tintul@mail.ru`, phone: +7(495)9393310, ORCID: https://orcid.org/0000-0002-5466-1221)

УДК 519.872:519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2021-29-3-260-270

Сдвинутые точки Соболя и многосеточный расчёт методом Монте-Карло

А. А. Белов[1, 2], **М. А. Тинтул**[1]

[1] *Московский государственный университет им. М. В. Ломоносова*
Ленинские горы, д. 1, стр. 2, Москва, 119991, Россия
[2] *Российский университет дружбы народов*
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия

Многомерные интегралы возникают во многих задачах физики. Например, моменты функции распределения в задачах переноса различных частиц (фотонов, нейтронов и др.) являются 6-мерными интегралами. При расчёте коэффициентов электропроводности и теплопроводности возникают интегралы рассеяния, размерность которых равна 12. Возникают задачи и с существенно большим числом переменных. Для вычисления интегралов столь высокой кратности наиболее эффективен метод Монте-Карло. Однако работоспособность этого метода сильно зависит от выбора последовательности, имитирующей набор случайных чисел. В литературе описано большое количество генераторов псевдослучайных чисел. Их качество проверяется с помощью батарей формальных тестов. Однако простейший визуальный анализ показывает, что прохождение таких тестов не гарантирует хорошей равномерности этих последовательностей. Для вычисления многомерных интегралов наиболее эффективны магические точки Соболя. В данной работе предложено усовершенствование этих последовательностей — смещённые магические точки Соболя, обеспечивающие большую равномерность распределения точек в многомерном кубе. Это ощутимо повышает точность кубатур. Существенной трудностью методов Монте-Карло является апостериорное подтверждение фактической точности. В данной работе предложен многосеточный алгоритм, позволяющий найти сеточное значение интеграла одновременно со статистически достоверной оценкой его точности. Ранее такие оценки были неизвестны. Проведены расчёты представительных тестовых интегралов с высокой фактической размерностью до 16. В качестве подынтегральной функции выбрана многомерная функция Вейерштрасса, не имеющая производной ни в одной точке. Эти расчёты убедительно показывают преимущества предложенных методов.

Ключевые слова: многомерный интеграл, метод Монте-Карло, точки Соболя, многосеточный расчет, апостериорные оценки точности

Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (online), 2658-4670 (print)

2021, 29 (3) 271–284
http://journals.rudn.ru/miph

Research article

UDC 519.872, 519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2021-29-3-271-284

Richardson–Kalitkin method in abstract description

Ali Baddour[1], **Mikhail D. Malykh**[1, 2]

[1] *Peoples' Friendship University of Russia (RUDN University)*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[2] *Meshcheryakov Laboratory of Information Technologies*
Joint Institute for Nuclear Research, Dubna, Russia
6, Joliot-Curie St., Dubna, Moscow Region, 141980, Russian Federation

(received: August 26, 2021; accepted: September 9, 2021)

An abstract description of the Richardson–Kalitkin method is given for obtaining a posteriori estimates for the proximity of the exact and found approximate solution of initial problems for ordinary differential equations (ODE). The problem $\mathcal{P}$ is considered, the solution of which results in a real number u. To solve this problem, a numerical method is used, that is, the set $H \subset \mathbb{R}$ and the mapping $u_h : H \to \mathbb{R}$ are given, the values of which can be calculated constructively. It is assumed that 0 is a limit point of the set H and u_h can be expanded in a convergent series in powers of h: $u_h = u + c_1 h^k + \dots$. In this very general situation, the Richardson–Kalitkin method is formulated for obtaining estimates for u and c from two values of u_h. The question of using a larger number of u_h values to obtain such estimates is considered. Examples are given to illustrate the theory. It is shown that the Richardson–Kalitkin approach can be successfully applied to problems that are solved not only by the finite difference method.

Key words and phrases: finite difference method, ordinary differential equations, a posteriori errors

1. Introduction

A priori estimates for finding solutions to dynamical systems using the finite difference method predict an exponential growth of the error with increasing time [1]. Therefore, long-term computation requires such a small sampling step that cannot be accepted in practice. Nevertheless, calculations for long times are carried out and it is generally accepted that they reproduce not the coordinates themselves, but some average characteristics of the trajectories. In this case, a posteriori error estimates are used instead of huge a priori ones. As early as in the works of Richardson [2], for estimating the errors arising in the calculation of definite integrals by the method of finite differences, it was proposed to refine the grid, and in the works of Runge a similar technique was applied to the study of ordinary differential equations. This approach was systematically developed in the works of N.N. Kalitkin and his disciples [3]–[7] as the Richardson method, although, given the role of Kalitkin in its development, it would be more correct to call it the Richardson–Kalitkin method.

The method itself is very general and universal, so we set out to present it in general form, divorcing it from the concrete implementation of the finite difference method. However, it soon became clear that this method could be extended to methods that are not finite difference methods, for example, the method of successive approximations, and even problems that are not related to differential equations.

In our opinion, this method is especially simply described for a class of problems in mechanics and mathematical physics, when it is necessary to calculate a significant number of auxiliary quantities, although only one value of some combination of them is interesting.

Example 1. On the segment $[0, T]$, we consider the initial problem

$$\frac{dx}{dt} = f(x,t), \quad x(0) = x_0,$$

it is required to find the value of x at the end of this segment, i.e., $x(T)$. To find this value numerically, we will have to calculate x approximately over the entire segment.

Example 2. On the segment $[0, T]$, we consider the dynamical system

$$\frac{dx}{dt} = f(x,y,t), \quad \frac{dy}{dt} = g(x,y,t),$$

with initial conditions $x(0) = x_0$, $y(0) = y_0$. It is required to find the value of the expression $x + y$ at the point $t = T$. To find this value numerically, we also have to calculate approximately x and y over the entire segment, then add the final values.

Example 3. The problem of many bodies is considered, say, the solar system, and it is required to find out whether the bodies scatter in 10 thousand years, or not. To solve it, it is enough to calculate the sum of the squares of the distances between the bodies and the center of mass of the system in 10 thousand years. At the same time, the coordinates and velocities of the bodies themselves are of no interest to anyone exactly 10 thousand years later.

Example 4. Let K be a unit circle on the plane. Find the first eigenvalue of the problem

$$\Delta v + \lambda v = 0, \quad v\big|_{\partial K} = 0.$$

Here, the eigenvalue λ_1 is to be found. We cannot find it numerically without finding the eigenfunction or roots of the determinant, i.e., other eigenvalues.

All these problems have one property in common: the result of the solution is a real number u. Various numerical methods are used to solve such problems. To substantiate these methods, the errors that occur in intermediate calculations when calculating auxiliary parameters are estimated, and then they are summed up. The a priori error estimates obtained in this way turn out to be enormous. However, in many cases the real situation is much better than the forecasts obtained in this way. Using example 2, this can be explained as follows: errors usually made in the calculation of x and y have different signs and therefore their contributions to the expression $x + y$ are canceled. Having estimated the error in calculating $x + y$ as the sum of the modules of errors in determining x and y, we inevitably and significantly overestimate the error. It will not be superfluous to note that problems whose solution is just a real number are considered in the topology $\mathbb{R}$. This means that the numerical solution must be a number that is close to the exact solution in that topology. However, the topology of the space in which the auxiliary variables take values is not specified. Usually, numerical methods are constructed so that these auxiliary variables are found with greater accuracy with respect to some Euclidean norm. For example, to find $x + y$ at time T, you need to find an approximation to the pair of functions $x(t)$, $y(t)$ with respect to the norm

$$\sup_{0 \leqslant t \leqslant T} \sqrt{|x(t)|^2 + |y(t)|^2}.$$

In the situation under consideration, such requirements are unnecessarily stringent.

In this paper, we describe a method for obtaining estimates of errors made in solving problems of this class in general form based on the Richardson–Kalitkin method [3], [4], abstracting from the particular choice of numerical method. In our opinion, this approach makes it possible to clearly see the main ideas of the Kalitkin method, which usually turn out to be hidden behind the details of the numerical methods used. Half a century of using the Richardson–Kalitkin method in practice has shown that its correct application requires the calculation of not two, but a significantly larger number of approximate solutions to test the hypothesis of the dominance of the principal term in the error (see section 4 below). We will discuss one possible modification of the method for the simultaneous use of all of these solutions for evaluating solutions and errors.

2. Basic definitions

Let the problem $\mathcal{P}$ be given, the solution of which is a real number u. We will not concretize this problem, let it only be known that this problem has a solution and, moreover, a unique one.

We are not going to concretize the numerical method for solving this problem. The use of any numerical method for solving it means replacing the problem $\mathcal{P}$ with another problem $\mathcal{P}_h$, the result of which is the mapping $u_h : H \to \mathbb{R}$. The interpretation of the set H essentially depends on the numerical method used. In some cases this set is a segment $(0, \infty)$, and in other cases it consists of positive rational numbers. For example, for the finite difference method, this set is formed by the admissible step lengths. Below this does not matter, but it is important that the set H is a subset of the real axis and that 0 is a limit point for the set H.

By analogy with the usual conventions, let us accept the following

Definition 1. Let $u_h : H \to \mathbb{R}$ be a solution to the problem $\mathcal{P}_h$. If $\lim\limits_{h\to 0} u_h = u$, then we say that the problem $\mathcal{P}_h$ approximates the problem $\mathcal{P}$. If $u_h = u + \mathcal{O}(h^k)$, then we say that the order of approximation of problem $\mathcal{P}$ by problem $\mathcal{P}_h$ is k.

In the overwhelming majority of cases, the value h has the meaning of the discretization step of the original problem, and the order of k is known. Here are some examples.

Example 5. Let the problem $\mathcal{P}$ consist in finding the value of the integral

$$u = \int\limits_{x=0}^{1} \frac{dx}{1+x^2}.$$

Its solution is the number $u = \pi/4$, which we do not know exactly. To calculate it, we cut the segment $[0,1]$ into $N \in \mathbb{N}$ parts. Let us assume that H is formed by all possible inverse natural numbers. Let u_h map this set to $\mathbb{R}$, putting in correspondence to $h = \frac{1}{N}$ the number

$$\sum_{n=0}^{N} \frac{h}{1+(nh)^2}.$$

Then $u_h = \pi/4 + \mathcal{O}(h)$, i.e., the order of approximation obtained by the rectangle rule is 1.

Remark 1. It should be noted that the methods of the numerical calculation of some classes integrals are known when the error depends on the step value not linearly or quadratically, but exponentially [8], [9].

Example 6. Let us consider the problem from example 1. An explicit Euler scheme can be used to solve it. We cut the segment $[0,T]$ into $N \in \mathbb{N}$ parts and take $h = \frac{T}{N}$. Let us put this number in correspondence with the number $u_h = x_N$, which is calculated by the recurrent formulas

$$x_{n+1} = x_n + f(x_n, nh)h, \quad n = 0, \ldots, N-1.$$

Moreover, it is possible to prove an a priori estimate for the error [1]:

$$|u + u - u_h| = |x(T) - x_N| \leqslant Ce^{aT}h,$$

where C, a are some constants depending only on f and the initial data x_0, but not on h and T. This immediately implies that $u_h = x(T) + \mathcal{O}(h)$, i.e., the order of approximation by the problem obtained using the Euler scheme is 1.

Basically, the finite difference method will be applied further, but this is not at all necessary.

Example 7. To calculate $u = x(T)$ from example 1, one can use the sequential iteration method (Picard's method). Let $N \in \mathbb{N}$ be the number of iterations, let us take $h = 1/N$ and assign this number to the number u_h, which is calculated as follows. First, N functions are calculated by recurrent formulas

$$x_{n+1}(t) = \int\limits_{\tau=0}^{t} f(x_n(\tau), \tau) d\tau, \quad n = 0, \dots, N-1,$$

and then $u_h = x_N(T)$. In this case $u_h \to u$ at $h \to 0$, i.e. the problem $\mathcal{P}_h$ approximates the initial problem.

The problem $\mathcal{P}_h$ should be simpler than the original one in the sense that it is possible to calculate the values of the mapping $u_h : H \to \mathbb{R}$ at all points of H. In practice, this possibility is limited both by an increase in the computational complexity when approaching $h = 0$, and by an increase in the role of the round-off error.

Definition 2. The value of the function u_h at any point of the set H will be called the approximate solution to the problem $\mathcal{P}$, and the modulus of the difference between this value and the solution to the problem $\mathcal{P}$ is the error made when solving problems $\mathcal{P}$ by method $\mathcal{P}_h$.

3. A posteriori error estimates

The Richardson–Kalitkin method can be separated from the finite difference method by adopting the following definition.

Definition 3. Let $u_h : H \to \mathbb{R}$ be a solution to the problem $\mathcal{P}_h$. If there exists a constant $c \neq 0$ such that $u_h = u + ch^k + \mathcal{O}\left(h^{k+1}\right)$, then we will say that ch^k is the leading term of the approximation error for problem $\mathcal{P}$ by problem $\mathcal{P}_h$.

Remark 2. In practice, it is usually assumed that the estimate $u_h = u + \mathcal{O}(h^k)$ implies the existence of a constant c such that $u_h = u + ch^k + \mathcal{O}(h^{k+1})$. Usually, this can be justified. But the definition 3 specifically states that $c \neq 0$. If $c = 0$, then one speaks of superconvergence of the method, because the order of approximation turns out to be greater than that predicted in theory. For difficulties in applying the Richardson–Kalitkin method in the case of superconvergence, see [10].

The essence of the Richardson–Kalitkin method is as follows. If we discard $\mathcal{O}(h^{k+1})$, then $u_h = u + ch^k$. We do not know the values of u and c, but we can calculate u_h for any value of h. Taking two such values, say h_1 and h_2, we have a system of two linear equations

$$u_h(h_1) = u + ch_1^k, \quad u_h(h_2) = u + ch_2^k,$$

resolving which for u and c, we will find some estimates for these quantities. We are talking about estimates, not values, since they are obtained by discarding $\mathcal{O}(h^{k+1})$.

Definition 4. Let $u_h : H \to \mathbb{R}$ be a solution to the problem $\mathscr{P}_h$ and there exists a constant $c \neq 0$ such that $u_h = u + ch^k + \mathcal{O}(h^{k+1})$. For any two $h_1, h_2 \in H$ the solution to the system

$$u_h(h_1) = u + ch_1^k, \quad u_h(h_2) = u + ch_2^k$$

with respect to u and c will be called the Richardson–Kalitkin estimate for the solution u to the problem $\mathscr{P}$ and the coefficient c at the leading term of the approximation error. We will denote these estimates as $\tilde{u}(h_1, h_2)$ and $\tilde{c}(h_1, h_2)$, below we will often omit the indication of their dependence on h_1, h_2, if this will not introduce ambiguity into presentation.

Example 8. Consider the initial problem

$$\dot{x} = -y, \quad \dot{y} = x, \quad x(0) = 1, \quad y(0) = 0,$$

and let it be required to find $u = x(1)$. We approximate it according to the explicit Euler scheme and calculate the approximate solution for $h_1 = 0.1$ and $h_2 = 0.01$ in Sage [11]:

$$u_h(h_1) = 0.5707904499, \quad u_h(h_2) = 0.543038634332351.$$

The solution of the system

$$u_h(h_1) = u + ch_1, \quad u_h(h_2) = u + ch_2$$

yields an estimate $\tilde{u} = 0.539955099269280$ for $u = \cos 1 = 0.540302305868140$, and for the coefficient of the leading term of the error $\tilde{c} = 0.308353506307201$.

The result looks very reasonable. With $h = 0.1$, we have an estimate for the error $\tilde{c}h = 0.0308$, while the error itself is 0.0304. With $h = 0.01$, we have an estimate for the error $\tilde{c}h = 0.00308$, while the error itself is 0.0027. The estimate for the solution differs from the solution by only $3.5 \cdot 10^{-4}$, which is an order of magnitude better than the result with the smallest step.

Richardson–Kalitkin estimates can also be performed in problems for the solution of which other numerical methods are used, while for specific methods such estimates themselves are well known, but under different names. For example, in this way the error is estimated when determining the eigenvalues by means of the finite element method (FEM) [12].

Example 9. Let it be required to find the smallest eigenvalue of the problem

$$\Delta v + \lambda v = 0, \quad v|_{\partial K} = 0$$

in the unit circle K. Then the answer is the number $u = \lambda_1$. Let us apply the FEM implementation in the system FreeFem++ [13]. The parameter h will be the value of $1/N$, where N is the number of points into which the circle is divided during triangulation. Then, when using linear elements, the smallest eigenvalue of the approximate problem is $u_h = u + ch^2 + \mathcal{O}(h^3)$.

Two-sided estimates for the error were obtained in the PhD thesis by Panin [14]. Let us take $h_1 = 1/20$ at random, and $h_2 = 1/100$, then

$$u_h(h_1) = 6.0173, \quad u_h(h_2) = 5.79292.$$

The solution of the system

$$u_h(h_1) = u + ch_1^2, \quad u_h(h_2) = u + ch_2^2$$

yields $\tilde{u} = 5.78357083333333$ against the exact value $u = j_1^2 = 5.783185962946785$, and for the coefficient of the leading term of the error we get $\tilde{c} = 93.4916666666667$.

The result looks very reasonable. For $h = 0.01$, we have an estimate for the error $\tilde{c}h^2 = 9.34 \cdot 10^{-3}$, while the error itself is $9.73 \cdot 10^{-3}$. The estimate for the solution differs from the solution by only $3.84 \cdot 10^{-4}$, which is an order of magnitude better than the result for the least h.

4. Justification of the Richardson–Kalitkin method

Justification of the Richardson–Kalitkin method consists of two parts: first, it is necessary to prove that the used numerical method satisfies the asymptotic formula

$$u_h = u + ch^k + \mathcal{O}(h^{k+1}).$$

Second, it is necessary to justify the possibility of omitting $\mathcal{O}(h^{k+1})$. The first step essentially depends on the numerical method used and its discussion is beyond the scope of this article. The second step, on the contrary, has nothing to do with the choice of a numerical method. Let us consider it in more detail. To discard the remainder $\mathcal{O}(h^{k+1})$, it must be substantially less than the principal term ch^k. For this purpose, first of all, c must be nonzero, which is indicated in definition 4. Further, the considered values of h should be sufficiently small. We have no a priori data to know in advance how small the chosen h should be. Finally, in practice, we cannot take h too small as well, when the round-off error becomes essential in the calculation of u_h.

In order to find a practically suitable interval of h values, N. N. Kalitkin and his disciples [5]–[7] have recommended to carry out calculations at least at 10 points rather than only two ones. Richardson's method can be applied only for those h, for which the error versus the step plotted in the log-log scale using these points, lies on a straight line with the slope k known from the theory. If the steps are too large, this plot differs from the straight line due to the fact that the discarded $\mathcal{O}(h^{k+1})$ is still large, and if the steps are too small, the rounding error becomes essential. If the slope of the straight line differs from k, then the phenomenon of superconvergence takes place (see remark 2).

Example 10. Let us return to example 8 and find an approximate solution by the fourth-order Runge–Kutta method with 15 steps, starting from the step $dt = 0.1$, each time decreasing the step by two times. Taking the approximate value for $x(1) = \cos 1$, obtained at the smallest step, as exact, we can plot the dependence of the error $\Delta x = x_n - x_{15}$ on the step dt, see the figure 1. The plot clearly shows an inclined section with a slope of approximately 4, followed by a horizontal section, interpreted as a region where a round-off error prevents further refinement of the solution.

With this approach, several natural questions arise. First, the points never exactly fall on a straight line. Therefore, we need quantitative characteristics for the site, which we will consider straight. How can we find them? Second, since approximate solutions were found not for two, but for many values of h, how can they be used to refine the solution? Third, the terms in power series do not have to form a monotonic sequence, therefore, for large h, the leading term can be significantly less than the next term. Can this possibility be taken into account explicitly?

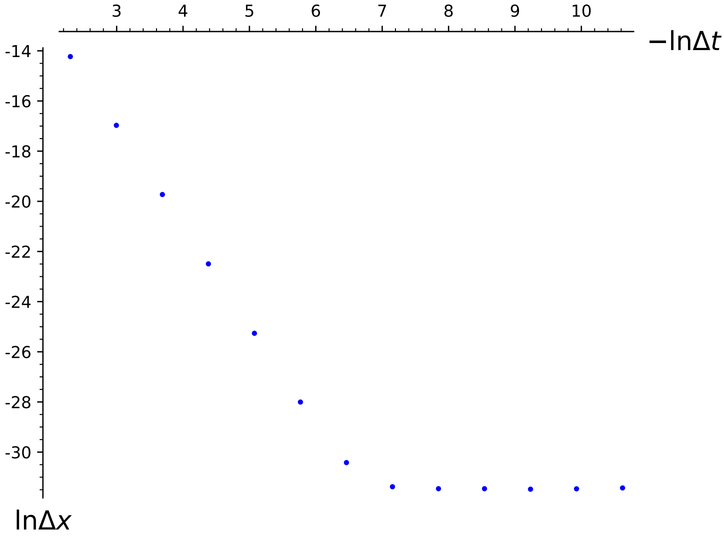


Figure 1. Dependence of error on step for example 10

5. Usage of several terms in the expansion of u_h in powers of h

The simplest answers to these questions can be found if we take into account the following terms in the expansion of u_h in powers of h. Suppose that u_h expands into a power series

$$u_h = u + c_1 h^k + c_2 h^{k+1} + \dots \tag{1}$$

If we have performed calculations for N different values for h, say, for $h = h_1, \dots, h_N$, then we can estimate the value of u and $N-1$ coefficients, discarding all terms, starting with $c_N h^N$.

Definition 5. Let the solution $u_h : H \to \mathbb{R}$ to the problem $\mathcal{P}_h$ be expanded in a power series (1), and let there be nonzero coefficients among $c_1, \dots c_{N-1}$. For any N values $h_1, \dots h_N \in H$ the solution to the system

$$u_h(h_n) = u + c_1 h_n^k + \dots + c_{N-1} h_n^{N+k-1}, \quad n = 1, 2, \dots, N \tag{2}$$

with respect to u and $c_1, \dots, c_{N-1}$ will be called an estimate for the solution u to the problem $\mathcal{P}$ and the first coefficients c over N approximate solutions. We will denote these estimates as $\tilde{u}$ and $\tilde{c}_1, \dots, \tilde{c}_{N-1}$.

As a result of solving system (2) we have: i) the estimate $\tilde{u}$ for the value of the exact solution, ii) the estimate $\tilde{c}_1 h^k$ for the error, suitable for sufficiently small h, and additional information about how small are those terms that are not taken into account in the Richardson–Kalitkin method.

Of course, as in the previous section, discarding terms, the order of which is equal to or greater than $N + k$ requires certain conditions to be met. However, these conditions are noticeably less restrictive. First, the simultaneous vanishing of several expansion coefficients seems incredible. Second, we can consider sufficiently large values of h for which the subsequent terms of the expansion are still noticeable.

6. Computer experiments

In our tests, we took $N = 4$ and $h_1 \in \mathbb{Q} \cap H$ at random, and the remaining h_2, h_3, h_4 were obtained by dividing h_1 by 2, 3 and 4. To avoid introducing additional rounding errors, system (2) is solved exactly over the field $\mathbb{Q}$.

Let us start with the simplest linear example.

Example 11. We will solve the problem from example 8 by the fourth-order Runge–Kutta method with a uniform step h. With step $h_1 = 0.1$, we get

$$u_h(0.1) = 0.540302967116884$$

against

$$\cos 1 = 0.540302305868140\ldots,$$

i.e., 6 correct decimal places. Calculating three more approximate solutions, we get the estimate for $u = \cos 1$ coinciding with the exact value up to 13 digits (the penultimate one). The estimate for the expansion coefficients (1) allows us to evaluate the error at $h = 0.1$ as

$$u_h - u = 0.007 \cdot 10^{-4} - 0.011 \cdot 10^{-5} + \ldots = 6 \cdot 10^{-7},$$

as it should be. It is interesting to compare the interpolation polynomials obtained at the initial step $dt = 0.1$ and $dt = 0.01$: the estimate for $u = \cos 1$ coincides with the one obtained earlier up to the last digit, c_1 differs in the fifth digit, c_2 differs by an order of magnitude, and c_3 — by two orders of magnitude. We increased the number of bits allocated to a real number, and made sure that the noted effects are not related to round-off errors.

In the course of our experiments, we came across situations where the coefficients are monstrously overestimated.

Example 12. Consider the same system

$$\dot{x} = -y, \quad \dot{y} = x, \quad x(0) = 1, \quad y(0) = 0,$$

but let it be required to find $u = x(0.3)$. At the first step, $h_1 = 10^{-4}$, we got a huge estimate $\tilde{c} = 5 \cdot 10^{13}$, while the scatter of estimates is very high. However, the estimate for $\cos 0.3$ itself coincides with the exact value with a very high accuracy, and one can easily find such values for the initial step, at which the estimates for the coefficients look quite reasonable.

Application of the standard Richardson–Kalitkin method ($N = 2$) leads to even less pleasant results in this example. Take $h_1 = 0.1$ and $h_2 = 0.05$ and estimate u and c_1 using the Richardson–Kalitkin method. Then the estimate for the error $u_h - u$ will be $\tilde{c}_1 0.1^4 = 10^{-10}$, which is much less than the actual error $u_h(0.1) - \cos(0.1)$, equal to $2 \cdot 10^{-2}$.

The simplest explanation for these effects is that in the series

$$u_h = \cos 0.3 + c_1 h^4 + c_2 h^5 + ...$$

the coefficient c_1 is very small, but the coefficient for some large power of h, on the contrary, is very large. Because of this, firstly, for small steps of the order of $h = 0.01$, we already have a value that coincides with the exact one, and, secondly, our estimates, which are based on the assumption of the possibility of discarding senior terms, do not work.

Now we proceed to a simplest nonlinear example.

Example 13. Let it be required to find $u = x(1)$ for solving the initial Volterra–Lotka problem

$$\dot{x} = (1 - y)x, \quad \dot{y} = -(1 - x)y, \quad x(0) = 0.5, \quad y(0) = 2$$

on the segment $0 < t < 1$. We will solve this problem according to the explicit Runge-Kutta scheme of the 4th order and estimate the solution with four steps, starting with $h_1 = 0.1$. For u, we obtain the estimate

$$u = 0.302408337777406,$$

and for the error

$$u_h(h) - u = -0.002 \cdot dt^4 + 0.00001 \cdot dt^5 + \dots .$$

At the smallest step, we have an error of 10^{-9}, that is, we can rely on more than 9 decimal places. Starting with $h_1 = 0.01$ we get another estimate, in which $\tilde{u}$ differs from the previously found value in the last two digits, and $\tilde{c}_1$ differs in the fourth digit.

7. Discussion of experimental results

The experiments performed, first of all, indicate that the proposed generalization of the Richardson–Kalitkin method allows, with a very modest number of steps, to obtain an estimate for the exact solution that coincides with it up to a round-off error. In this case, instead of 1 calculation, we perform 4 independent ones, which does not waste time at all, since the calculations are performed in parallel.

The larger the power, the greater the discrepancy in determining the coefficients for powers of h. It is not hard to explain this fact. All formulas are derived under an assumption typical of various kinds of mean-value theorems: for any $s > n$ there is a constant M_s such that

$$\left\| x - c_0 - \sum_{i=1}^{s} c_i h^{n+i-1} \right\| \leqslant M_s h^{n+s+1}.$$

When solving the interpolation problem, we solve the problem

$$c_0 + \sum_{i=1}^{s} c_i h_j^{n+i-1} = b_j + \xi_j h_j^{n+s+1},$$

where b_j are the values of x for $h = h_j$, and ξ_j are unknown quantities, about which we know that $|\xi_j| \leqslant M_s$.

Consider, for simplicity, $s = 2$

$$c_0 + c_1 h_1^n = b_1 + \xi_1 h_1^{n+1}, \quad c_0 + c_1 h_2^n = b_2 + \xi_1 h_2^{n+1}.$$

According to Cramer's formulas

$$c_0 = \frac{b_2 h_1^n - b_1 h_2^n}{h_1^n - h_2^n} - (h_1 h_2)^n \frac{h_1 \xi_1 - h_2 \xi_2}{h_1^n - h_2^n},$$

and

$$c_1 = \frac{b_1 - b_2}{h_1^n - h_2^n} + \frac{h_1^{n+1}\xi_1 - h_2^{n+1}\xi_2}{h_1^n - h_2^n}.$$

For $h_1 = h$, $h_2 = h/2$, the error in c_0 will be of the order of $O(h^{n+1})$, and in c_1 — only of the order of $O(h)$. As s grows, the divergence of orders will become more and more noticeable.

Of course, the main problem is that we do not know neither s nor M_s. The example, in which superconvergence manifested itself, makes one think that there are cases when s cannot be taken as wanted. But in this case, the problem of applicability of the described method is reduced to the classical problem of the theory of power series: how many terms should be taken in the series in order to have a given accuracy? It is not difficult to answer it if the recurrent formulas for the coefficients are known, rather than the estimates for the coefficients of the power series, which become the worse the greater the power.

In theory, this circumstance is obviously a serious problem. However, in fact, all problematic cases immediately manifested themselves in the form of inadequately large coefficients. Thus, as a practical recipe, the generalization described seems to be quite useful.

8. Conclusion

We described the Richardson–Kalitkin method as a means for evaluating numerical methods for solving any problem $\mathcal{P}$, the result of which is a real number u. To specify a numerical method for solving the problem $\mathcal{P}$ means to specify the set $H \subset \mathbb{R}$, for which 0 is a limit point, and the mapping $u_h : H \to \mathbb{R}$, the values of which can be calculated constructively. This method gives a solution to the problem $\mathcal{P}$, if $\lim\limits_{h\to 0} u_h = u$.

If there exist $k \in \mathbb{N}$ and numbers $c_1, \dots, c_N$, among which there are nonzero numbers such that

$$u_h = u + ch^k + \dots c_N h^{k+N} + \mathcal{O}(h^{k+N+1}),$$

then from N values of the mapping u_h it is possible to estimate the exact solution of the original problem and the coefficients $c_1, \dots, c_N$, characterizing the error of the numerical method. The examples show that the higher the coefficient number, the worse these estimates are, but on the whole they characterize the numerical method quite accurately. The values of u_h are calculated independently, so the calculation of such problems can be naturally parallelized.

Acknowledgments

This work is supported by the Russian Science Foundation (grant no. 20-11-20257).

References

[1] E. Hairer, G. Wanner, and S. P. Nørsett, *Solving Ordinary Differential Equations*, 3rd ed. New York: Springer, 2008, vol. 1.

[2] L. F. Richardson and J. A. Gaunt, “The deferred approach to the limit,” *Phil. Trans. A*, vol. 226, pp. 299–349, 1927. DOI: 10.1098/rsta.1927.0008.

[3] N. N. Kalitkin, A. B. Al’shin, E. A. Al’shina, and B. V. Rogov, *Calculations on quasi-uniform grids*. Moscow: Fizmatlit, 2005, In Russian.

[4] N. N. Kalitkin, *Numerical methods [Chislennyye metody]*. Moscow: Nauka, 1979, In Russian.

[5] A. A. Belov, N. N. Kalitkin, and I. P. Poshivaylo, “Geometrically adaptive grids for stiff Cauchy problems,” *Doklady Mathematics*, vol. 93, no. 1, pp. 112–116, 2016. DOI: 10.1134/S1064562416010129.

[6] A. A. Belov and N. N. Kalitkin, “Nonlinearity problem in the numerical solution of superstiff Cauchy problems,” *Mathematical Models and Computer Simulations*, vol. 8, no. 6, pp. 638–650, 2016. DOI: 10.1134/S2070048216060065.

[7] A. A. Belov, N. N. Kalitkin, P. E. Bulatov, and E. K. Zholkovskii, “Explicit methods for integrating stiff Cauchy problems,” *Doklady Mathematics*, vol. 99, no. 2, pp. 230–234, 2019. DOI: 10 . 1134 / S1064562419020273.

[8] L. N. Trefethen and J. A. C. Weideman, “The exponentially convergent trapezoidal rule,” *SIAM Review*, vol. 56, pp. 385–458, 3 2014. DOI: 10.1137/130932132.

[9] A. A. Belov and V. S. Khokhlachev, “Asymptotically accurate error estimates of exponential convergence for the trapezoid rule,” *Discrete and Continuous Models and Applied Computational Science*, vol. 3, pp. 251–259, 2021. DOI: 10.22363/2658-4670-2021-29-3-251-259.

[10] A. Baddour, M. D. Malykh, A. A. Panin, and L. A. Sevastianov, "Numerical determination of the singularity order of a system of differential equations," *Discrete and Continuous Models and Applied Computational Science*, vol. 28, no. 5, pp. 17–34, 2020. DOI: 10.22363/2658-4670-2020-28-1-17-34.

[11] The Sage Developers. "SageMath, the Sage Mathematics Software System (Version 7.4)." (2016), [Online]. Available: https://www.sagemath.org.

[12] O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu, *The finite element method: its basis and fundamentals*, 7th ed. Elsiver, 2013.

[13] F. Hecht, "New development in FreeFem++," *Journal of Numerical Mathematics*, vol. 20, no. 3–4, pp. 251–265, 2012. DOI: 10.1515/jnum-2012-0013.

[14] A. A. Panin, "Estimates of the accuracy of approximate solutions and their application in the problems of mathematical theory of waveguides [Otsenki tochnosti priblizhonnykh resheniy i ikh primeneniye v zadachakh matematicheskoy teorii volnovodov]," in Russian, Ph.D. dissertation, MSU, Moscow, 2009.

For citation:

A. Baddour, M. D. Malykh, Richardson–Kalitkin method in abstract description, Discrete and Continuous Models and Applied Computational Science 29 (3) (2021) 271–284. DOI: 10.22363/2658-4670-2021-29-3-271-284.

Information about the authors:

Baddour, Ali — PhD student of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: alibddour@gmail.com, phone: +7(495)9550927, ORCID: https://orcid.org/0000-0001-8950-1781)

Malykh, Mikhail D. — Doctor of Physical and Mathematical Sciences, Assistant professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University); Researcher in Meshcheryakov Laboratory of Information Technologies, Joint Institute for Nuclear Research (e-mail: malykh_md@pfur.ru, phone: +7(495)9550927, ORCID: https://orcid.org/0000-0001-6541-6603, ResearcherID: P-8123-2016, Scopus Author ID: 6602318510)

УДК 519.872, 519.217
PACS 07.05.Tp, 02.60.Pn, 02.70.Bf
DOI: 10.22363/2658-4670-2021-29-3-271-284

Метод Ричардсона–Калиткина в абстрактном изложении

Али Баддур[1], **М. Д. Малых**[1, 2]

[1] *Российский университет дружбы народов*
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия
[2] *Лаборатория информационных технологий им. М. Г. Мещерякова*
Объединённый институт ядерных исследований
ул. Жолио-Кюри, д. 6, Дубна, Московская область, 141980, Россия

Дано абстрактное описание метода Ричардсона-Калиткина для получения апостериорных оценок близости точного и найденного приближённого решения начальных задач для обыкновенных дифференциальных уравнений (ОДУ). Рассматривается задача $\mathcal{P}$, результатом решения которой является вещественное число u. Для решения этой задачи используется численный метод, то есть заданы множество $H \subset \mathbb{R}$ и отображение $u_h : H \to \mathbb{R}$, значения которого имеется возможность вычислять конструктивно. При этом предполагается, что 0 является предельной точкой множества H, u_h можно разложить в сходящийся ряд по степеням h: $u_h = u + c_1 h^k + \ldots$. В этой весьма общей ситуации сформулирован метод Ричардсона–Калиткина получения оценок для u и c по двум значениям u_h. Рассмотрен вопрос об использовании большего числа значений u_h для получения такого рода оценок. Приведены примеры, иллюстрирующие теорию. Показано, что подход Ричардсона–Калиткина с успехом может быть применён к задачам, которые решаются не только методом конечных разностей.

Ключевые слова: метод конечных разностей, обыкновенные дифференциальные уравнения, апостериорные ошибки