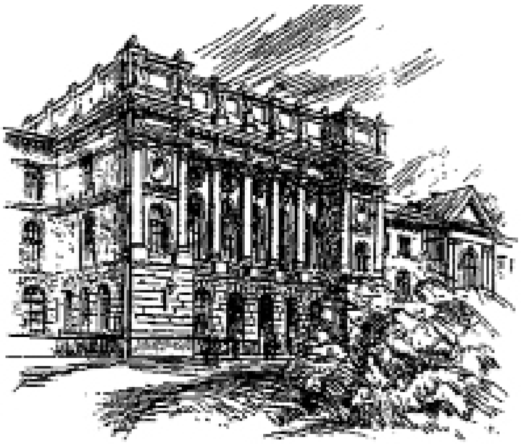
THE MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE RUSSIAN FEDERATION



ISSN 2687-0517

Computing, Telecommunications and Control

Vol. 18, No. 2
2025

Peter the Great St. Petersburg
Polytechnic University
2025

COMPUTING, TELECOMMUNICATIONS AND CONTROL

The journal is included in the List of Leading PeerReviewed Scientific Journals and other editions to publish major findings of PhD theses for the research degrees of Doctor of Sciences and Candidate of Sciences.

Open access journal is to publish articles of a high scientific level covering advanced experience, research results, theoretical and practical problems of informatics, electronics, telecommunications, and control.

The journal is indexed by Ulrich's Periodicals Directory, Google Scholar, EBSCO, ProQuest, Index Copernicus, VINITI RAS Abstract Journal (Referativnyi Zhurnal), VINITI RAS Scientific and Technical Literature Collection, Russian Science Citation Index (RSCI) database Scientific Electronic Library and Math-Net.ru databases.

The journal is registered with the Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications (ROSKOMNADZOR). Certificate ЭЛ No. ФС77-77378 issued 25.12.2019.

Editorial office

Dr. Sc., Professor A.S. Korotkov – Editor-in-Chief;

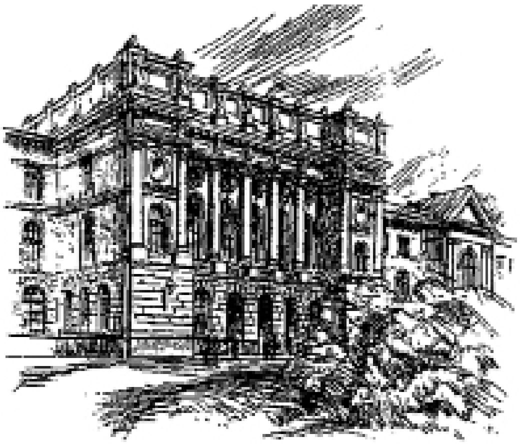
Ph.Ch.S. Bastian – literary editor, proofreader; G.A. Pyshkina – editorial manager; A.A. Kononova – computer layout; I.E. Lebedeva – English translation.

Address: 195251 Polytekhnicheskaya Str. 29, St. Petersburg, Russia.

+7 (812) 552-6216, e-mail: infocom@spbstu.ru

Release date: 06.06.2025

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ



ISSN 2687-0517

Информатика, телекоммуникации и управление

Том 18, № 2
2025

Санкт-Петербургский политехнический
университет Петра Великого
2025

ИНФОРМАТИКА, ТЕЛЕКОММУНИКАЦИИ И УПРАВЛЕНИЕ

Журнал с 2002 года входит в Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени доктора и кандидата наук.

Сетевое издание открытого доступа публикует статьи высокого научного уровня, освещающие передовой опыт, результаты НИР, теоретические и практические проблемы информатики, электроники, телекоммуникаций, управления.

Сведения о публикациях представлены в Реферативном журнале ВИНИТИ РАН, в международной справочной системе «Ulrich`s Periodical Directory», в Российской государственной библиотеке. В базах данных: Российский индекс научного цитирования (РИНЦ), Google Scholar, EBSCO, Math-Net.Ru, ProQuest, Index Copernicus.

Журнал зарегистрирован Федеральной службой по надзору в сфере информационных технологий и массовых коммуникаций (Роскомнадзор). Свидетельство о регистрации ЭЛ № ФС77-77378 от 25.12.2019.

Учредитель и издатель: Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация.

Редакция журнала

д-р техн. наук, профессор А.С. Коротков – главный редактор;

Ф.К.С. Бастиан – литературный редактор, корректор; Г.А. Пышкина – ответственный секретарь, выпускающий редактор;

А.А. Кононова – компьютерная вёрстка; И.Е. Лебедева – перевод на английский язык.

Адрес редакции: Россия, 195251, Санкт-Петербург, ул. Политехническая, д. 29.

Тел. редакции +7(812) 552-62-16, e-mail редакции: infocom@spbstu.ru

Дата выхода: 06.06.2025

Contents

Intelligent Systems and Technologies, Artificial Intelligence

Circuits and Systems for Receiving, Transmitting, and Signal Processing

Software and Hardware of Computer, Network, Telecommunication, Control, and Measurement Systems

Содержание

Интеллектуальные системы и технологии, искусственный интеллект

Устройства и системы передачи, приема и обработки сигналов

Компьютерные сети, вычислительные, телекоммуникационные, управляющие и измерительные системы

Research article
DOI: https://doi.org/10.18721/JCSTCS.18201
UDC 519.2

CATEGORICAL SURVIVAL ANALYSIS OF THE REQUIRED JOB EXECUTION TIMES IN THE HYBRID SUPERCOMPUTER CENTER

T.A. Misharina[1], *S.V. Malov*[1,2] ✉

[1] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation;

[2] St. Petersburg Electrotechnical University,
St. Petersburg, Russian Federation

✉ tanechkamisharina254@gmail.com

Abstract. According to statistics, the actual execution time of most jobs on a supercomputer cluster differs significantly from the time requested by the user. Investigation of distributions of supercomputer job execution times using statistical or machine learning methods allows optimizing the operation of a supercomputer cluster. We study the results of computational jobs processing in the supercomputer center of Peter the Great St. Petersburg Polytechnic University. We have developed a nonparametric approach for detection and statistical confirmation of weak stochastic orders based on categorical nonparametric framework of contrasts obtained from the Kaplan–Meier estimators obtained from independent groups of right-censored observations. To adjust the confidence level of the detected weak stochastic orders, we apply the Bonferroni correction to all the comparisons under consideration. We perform comparative statistical analysis of the distributions of required execution times to complete successfully the job in different groups of right-censored observations; detect and confirm available weak stochastic orders.

Keywords: survival data, Kaplan–Meier estimator, Wald's type test, stochastic orders, supercomputer cluster, job scheduling

Acknowledgements: The research was partially financially supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state assignment "Development and research of machine learning models for solving fundamental problems of artificial intelligence in the fuel and energy complex" (FSEG-2024-0027). The results of the work were obtained using the computational resources of the shared-use center "Polytechnic Supercomputer Center" of Peter the Great St. Petersburg Polytechnic University (No. 500675, https://ckp-rf.ru/catalog/ckp/500675/).

Citation: Misharina T.A., Malov S.V. Categorical survival analysis of the required job execution times in the hybrid supercomputer center. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 7–20. DOI: 10.18721/JCSTCS.18201

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18201
УДК 519.2

КАТЕГОРИАЛЬНЫЙ АНАЛИЗ ВЫЖИВАЕМОСТИ ТРЕБУЕМЫХ ВРЕМЕН ИСПОЛНЕНИЯ ЗАДАЧ В ГИБРИДНОМ СУПЕРКОМПЬЮТЕРНОМ ЦЕНТРЕ

Т.А. Мишарина[1], *С.В. Малов*[1,2] ✉

[1] Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация;

[2] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»
им. В.И. Ульянова (Ленина), Санкт-Петербург, Российская Федерация;

✉ tanechkamisharina254@gmail.com

Аннотация. Согласно статистике, фактическое время выполнения большинства заданий на суперкомпьютерном кластере существенно отличается от времени, запрошенного пользователем. Исследование распределений времен исполнения задач на суперкомпьютере с использованием статистических методов или методов машинного обучения позволяет оптимизировать работу суперкомпьютерного кластера. Мы изучаем результаты исполнения вычислительных задач в суперкомпьютерном центре Санкт-Петербургского Политехнического университета Петра Великого. Нами разработан непараметрический подход для обнаружения и подтверждения статистической достоверности слабых стохастических порядков. Данный подход основан на категориальном непараметрическом методе сравнений на базе оценок Каплана–Мейера, построенных по независимым группам цензурированных справа наблюдений. Для корректировки уровня достоверности обнаруженных слабых стохастических порядков мы применяем поправку Бонферрони на все рассматриваемые сравнения. Проведен сравнительный статистический анализ распределений времен, необходимых для корректного завершения задач, в различных группах наблюдений, найдены и статистически подтверждены некоторые слабые стохастические порядки.

Ключевые слова: данные типа времени жизни, оценка Каплана–Мейера, критерий типа Вальда, стохастические порядки, суперкомпьютерный кластер, планировщик задач

Финансирование: Исследование выполнено при частичной финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках государственного задания «Разработка и исследование моделей машинного обучения для решения фундаментальных задач искусственного интеллекта в топливно-энергетическом комплексе» (FSEG-2024-0027). Результаты работы получены с использованием вычислительных ресурсов центра коллективного пользования «Политехнический суперкомпьютерный центр» Санкт-Петербургского политехнического университета Петра Великого (№ 500675, https://ckp-rf.ru/catalog/ckp/500675/).

Для цитирования: Misharina T.A., Malov S.V. Categorical survival analysis of the required job execution times in the hybrid supercomputer center // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 7–20. DOI: 10.18721/JCSTCS.18201

Introduction

High performance computing is becoming increasingly important in different areas of scientific research and industry. Collective supercomputer centers allow to perform calculations of any complexity to a wide range of users. The operation of a supercomputer center is a complex parallel queuing process of execution of computational jobs over time. An optimal scheduling of entire jobs leads to increasing performance of computations. The most important characteristic of a job is the required supercomputer resource involving the required time for its execution, the number of cores allocated to provide sufficient

Random-Access Memory (RAM) and the job execution quickness. A job scheduled on supercomputer can be divided into computational tasks that can be executed in parallel on different cores. An exit code is obtained at the end of the execution of each job.

The optimization of jobs management systems was discussed by a number of authors. The most famous subject of interest is the job running time and its prediction time given by user. In most cases user overestimate significantly job running time that implies non optimality in job scheduling. A machine learning regression-based method to predict mean running time by a vector of observed futures was studied in [5]. It was shown that the prediction of job running time allows the correction of running times obtained from users that increases sufficiently efficiency of job scheduling. Applications of supervised machine learning algorithms to predict the job running time based on information submitted by user at high performance computing centers was discussed in [17]. Machine learning classification methods to predict a class of running time distribution was under consideration in [4, 18]. The underestimation effect of running time by user was studied in [6]. Note that the most important characteristic of job processing is the required execution time to complete successfully the job, which can be equal to the running time or not available, if the job is terminated. Using the observed running time instead of the required execution time as well as just removing jobs, which was not completed successfully, lead to sufficient underestimation of the required execution time if the number of "unsuccessful" jobs is valuable in compare with the total number of jobs. The right-censored survival data model, which is also applicable in the reliability theory, allows to estimate correctly the distribution of required execution time by using the running time and the indicator, which displays, if the job is completed successfully. Machine learning algorithms to predict the distribution of required execution time and its characteristics based on semiparametric and nonparametric regression models of survival analysis were studied in [14, 21].

Note that machine learning methods are more flexible in compare with statistical ones. The statistical conclusions are restricted to the probabilistic model of the experiment, but the statistical conclusions yield another kind of reliability of obtained results.

Categorical methods for survival right-censored data analysis are widely presented in the literature. In [20], likelihood ratio test for right-censored grouped survival data was studied and the chi-square limit distribution of the likelihood ratio test statistic was obtained. In [10], the chi-square test and the Wald's type test for right-censored survival data was obtained and the comparative analysis of the tests under Pitman alternatives was performed. A parametric Pearson's type test for right-censored survival was studied in [8]. In [1], modified versions of goodness of fit chi-square tests for simple and composite parametric null hypotheses in the nonparametric survival data models under presence or absence of the right censoring were obtained. Presence of one extra degree of freedom of the limit distribution in compare with the classical version of the chi-square test is noted and some examples are given. In [12], another approach was used to obtain chi-square test for complex parametric null hypothesis and the comparative analysis in Pitman's efficiency of the test with another version of chi-square test obtained in [1] was performed. An adaptive version of the chi-square test obtained in [10] with a random data-based choice of grouping intervals is given in [2]. The chi-square tests for testing the null hypothesis that the failure time distributions agree with some known parametric model for hazard rates was given in [9], and the chi-square test for agreement of the failure time distributions with some known semiparametric regression model (e.g., the semiparametric accelerated failure time model) in general case under time dependent covariate was obtained in [3]. Wald-type categorical tests for testing homogeneity null hypotheses in the nonparametric right-censored survival data model used in this work was studied in [15], which is universal and can be more efficient than the linear rank tests commonly used in nonparametric survival analysis under some alternatives.

Moreover, machine learning methods also use for survival right-censored data analysis. Random forest-based algorithms were used to analyze right-censored survival data in [7, 11]. In [19], the authors propose a new Transformer-based survival model, which estimates the patient-specific survival distribution. Another example of the application of machine learning methods to the analysis of randomly censored survival data is presented in [13]. Here, the authors propose a method based on the Beran estimator using neural kernels to estimate the conditional average treatment effect.

In this work, we study the results of users' jobs processing in the supercomputer center of Peter the Great St. Petersburg Polytechnic University. Since the limitations for the running time and the resource of supercomputer used to execute a job should be determined in advance, it is important to evaluate distributions of main characteristics of a job, which cannot be predicted exactly. We are interested in two important characteristics of the required resource: the execution time required to complete successfully the job in seconds, without taking into account the number of cores allocated, and the required computer execution time that is obtained by multiplying the required execution time by the number of cores allocated. We investigate the distributions of the required execution times and required computer times and perform a comparative analysis of the distributions in different groups of users.

Each observation contains the supercomputer resource used to perform the job: the job processing time (observed execution time), the number of cores and the amount of memory allocated and the exit code, which indicates whether the job was completed successfully or it was interrupted due to an error, user request, lack of memory or time allocated for the job execution. In the latter cases, the job is incomplete and the job execution time is assumed to be censored. We relate the execution time (or computer time) required to complete successfully the user's job with the failure time in right-censored survival data model. Then the job execution time and the indicator of successful completion of user's job, which is determined by the exit code, is the right-censored observation.

We apply categorical nonparametric statistical framework based on contrasts obtained from the Kaplan—Meier estimators in d independent groups of right-censored observations to obtain advanced statistical conclusions on distributions of failure times in different groups of observations. Let T be the job execution time or computer time and U be the censoring time. Each observation (X, δ) contains the job processing time $X = \min(T, U)$ and the indicator $\delta = \mathrm{I}_{\{T \le U\}}$, that is equal to 1, if the exit code indicates that the computational task is completed successfully and 0 otherwise. We study the distribution of T and its dependence on the grouping factor, which reflects the user's area of expertise. We create advanced categorical methods for right-censored survival data and apply them to perform comparative analysis of the distributions of job execution times and computer times T in 11 groups of user's domain of scientific expertise [16].

Wald's type categorical tests for survival data

The main object of statistical analysis is the distribution of the required execution time (or computer time) T. The required execution time T is not observed exactly, if the corresponding job is not completed successfully, in this case, we explain the true execution time of the job as an independent censoring time. A single observation consists of the true execution time $X = \min(T, U)$ and the binary job exit code $\delta = \mathrm{I}_{\{T \le U\}}$, which indicates whether the job was completed successfully or censored. We allow the distributions of the required execution times to differ in different groups of jobs and use a categorical covariate $z \in \{1, ..., d\}$ for grouping the data. The observed data contain the true execution times $X_i = \min(T_i, U_i)$ and the binary exit codes $\delta_i = \mathrm{I}_{\{T_i \le U_i\}}$, where T_i is the required execution time of i-th job, U_i is the independent random censoring time, and the covariate $z_i \in \{1, ..., d\}$, which determines the group, to which a corresponding observation belongs, $i = 1, 2, ..., n$. Let $S_z(t) = \mathbb{P}_z(T > t)$ be a completely unknown survival function of the required execution time in group z, $z = 1, 2, ..., d$. The homogeneity null hypothesis is as follows:

$$H_0 : S_1(t) = S_2(t) = \cdots = S_d(t), \quad t \in (-\infty, \infty).$$

We consider a weaker version of the null hypothesis, which requires the equalities of the survival functions S_j at some fixed points:

$$H_0^*: S_1\left(\vec{t}\right) = S_2\left(\vec{t}\right) = \cdots = S_d\left(\vec{t}\right), \ \ \vec{t} = \left(t_1, \ \ldots, \ t_k\right)^T.$$

Let $\hat{S}_z$ be the Kaplan–Meier estimator of the survival function S_z in different groups, $z = 1, 2, \ldots, d$. The asymptotic properties of the Kaplan–Meier estimators imply weak convergence

$$\sqrt{n_z}\left(\hat{S}_z\left(\vec{t}\right) - S_z\left(\vec{t}\right)\right) \Rightarrow N\left(0, \Sigma_j\right), \ \ z = 1, \ 2, \ \ldots, \ d$$

for any fixed vector of time points , where $N\ (0,\ \Sigma_j)$ is the mean zero Gaussian distribution with the matrix of covariance Σ_z, n_z is the number of observations in group z.

The matrix of covariance Σ_z has the following form:

$$\Sigma_z = \left(\sigma_{vu}^{(z)}\right)_{v=1,u=1}^{k,k},$$

where $\sigma_{vu}^{(z)}$ is the element of the limit covariance matrix of the values of the Kaplan–Meier estimator at time points t_v and t_u, $v, u = 1, \ldots, k$.

The term $\sigma_{vu}^{(z)}$ of the covariance matrix can be estimated by the following Greenwood formula:

$$\hat{\sigma}_{vu}^{(z)} = n_z \hat{S}_z\left(t_v\right)\hat{S}_z\left(t_u\right)\sum\nolimits_{l=1}^{\min(v,u)} \frac{D_l}{Y_l^*\left(Y_l^* - D_l\right)}, \ \ v,u = 1, \ \ldots, \ k,$$

where D_l is the number of jobs completed successfully at T_l, Y_l^* is the number of jobs not completed and not censored before T_l.

Then the estimate of the covariance matrix Σ_z has the following form:

$$\hat{\Sigma}_z = \left(\hat{\sigma}_{vu}^{(z)}\right)_{v=1,u=1}^{k,k}.$$

Taking into account the independence of observations in different groups, we obtain the following joint weak convergence:

$$\sqrt{n}\left(\hat{S}^*\left(\vec{t}\right) - \hat{S}^*\left(\vec{t}\right)\right) \Rightarrow N\left(0, D\Sigma^* D\right),$$

where $\hat{S}^* = \left(\hat{S}_1, \ \hat{S}_2, \ \ldots, \ \hat{S}_d\right)$ is the Kaplan–Meier estimator of the vector of survival functions $S^* =$ $= \left(S_1, \ S_2, \ \ldots, \ S_d\right)^T$;

$$\Sigma^* = \begin{pmatrix} \Sigma_1 & \mathrm{O} & \ldots & \mathrm{O} \\ \mathrm{O} & \Sigma_2 & \ldots & \mathrm{O} \\ \ldots & \ldots & \ddots & \ldots \\ \mathrm{O} & \mathrm{O} & \ldots & \Sigma_d \end{pmatrix},$$

$D = \mathrm{diag}(n^*)$ is the normalizing diagonal matrix with the elements of the vector $n^* = \left(n_1^*, \ \ldots, \ n_d^*\right): n_z^* =$ $= \sqrt{\left(n_z, \ n_z, \ \ldots, \ n_z\right)/n}, \ \ z = 1, \ \ldots, \ d$, at the diagonal; $D\Sigma^*D$ is the limit covariance matrix; O is the matrix of zeroes of size $k \times k$.

Denote $S^*\left(\vec{t}\right)=\theta^*=\left(\theta_1,\ \theta_2,\ \ldots,\ \theta_d\right)^T$ and $\theta_z=\left(\theta_{z1},\ \theta_{z2},\ \ldots,\ \theta_{zk}\right)^T$, $z = 1, ..., d$. Then the null hypothesis can be written as follows:

$$H_0^*:\theta_{11}=\theta_{21}=\cdots=\theta_{d1},\ \ldots,\ \theta_{1k}=\theta_{2k}=\cdots=\theta_{dk}.$$

Let $\psi = C^T\theta^*$, where C^T is the matrix of contrasts of size $q \times m$, $\theta^* = (\theta_1, \theta_2, ..., \theta_m)^T$, is the vector of size $m \times 1$. The contrasts matrix $\psi = (\psi_1, ..., \psi_q)^T$ contains linear functions of parameter ψ_j such that:

$$\psi_j=c_{1j}\theta_1+c_{2j}\theta_2+\cdots+c_{mj}\theta_m,\ \sum\nolimits_{i=1}^m c_{ij}=0,$$

where c_{ij} are the elements of the matrix C, $i = 1, 2, ..., m$, $j = 1, 2, ..., q$.

We use the following pairwise contrasts:

$$\begin{gathered}\psi_1=S_1\left(t_1\right)-S_2\left(t_1\right);\\ \cdots\\ \psi_k=S_1\left(t_k\right)-S_2\left(t_k\right);\\ \psi_{k+1}=S_2\left(t_1\right)-S_3\left(t_1\right);\\ \cdots\\ \psi_{2k}=S_2\left(t_k\right)-S_3\left(t_k\right);\\ \vdots\\ \psi_{(d-1)k}=S_{d-1}\left(t_k\right)-S_d\left(t_k\right).\end{gathered}$$

In terms of the parameters $\theta_{ij} = S_i(t_j)$, $i = 1, 2, ..., d$, $j = 1, 2, ..., k$, the vector function of contrasts ψ can be rewritten in the following matrix form:

$$\psi=C^T\theta^*=\begin{pmatrix}E & -E & O & \cdots & O & O\\ O & E & -E & & O & O\\ & \vdots & & \ddots & & \vdots\\ O & O & O & \cdots & E & -E\end{pmatrix}\cdot\begin{pmatrix}\theta_{11}\\ \vdots\\ \theta_{1k}\\ \theta_{21}\\ \vdots\\ \theta_{2k}\\ \vdots\\ \theta_{d1}\\ \vdots\\ \theta_{dk}\end{pmatrix},$$

where E is the identity matrix of $k \times k$, O is the $k \times k$-matrix of zeroes, the matrix of contrasts C^T is of size $kd \times k(d - 1)$, the parameter θ^* is of size $kd \times 1$ and the contrasts vector ψ is of size $k(d - 1) \times 1$.

In terms of the contrasts ψ the null hypothesis can be written as follows:

$$H_0^*:\psi_1=\psi_2=\cdots=\psi_{k(d-1)}=0.$$

The asymptotic normality of the estimators $\hat{\theta}^* = \hat{S}^*\left(\vec{t}\right)$ implies immediately the asymptotic normality of the estimators $\hat{\psi}$ of the corresponding contrasts ψ:

$$\sqrt{n}\left(\hat{\psi} - \psi\right) \Rightarrow N\left(0, C^T D \sum\nolimits^* DC\right). \qquad (1)$$

Let $\Gamma_\psi = C^T D \sum^* DC$ and $\hat{\Gamma}_\psi = C^T D \hat{\sum}^* DC$, where $\hat{\sum}^*$ is a consistent estimate of the asymptotic variance of the estimator $\hat{\psi}$ under the null hypothesis. The Wald's type test statistic for testing H_0^*.

$$n\hat{\psi}^T \hat{\Gamma}_\psi^{-1} \hat{\psi} \Rightarrow \chi_q^2,$$

has asymptotical χ_q^2-distribution under the null hypothesis. Under the fixed alternative $H_A^* : \psi = \psi_0$ the Wald's type test statistic has an asymptotical non-central $\chi_{\mu,q}^2$-distribution with the non-centrality parameter

$$\mu = n\psi_0^T \Gamma_\psi^{-1} \psi_0.$$

Detection and testing significance of stochastic orders

Let X and Y be random variables. The random variable X is stochastically less than the random variable $Y\left(X \le^{st} Y\right)$, if

$$F_X\left(t\right) \ge F_Y\left(t\right) \quad \text{for all} \quad t \in \left(-\infty, \infty\right),$$

where $F_X(t) = P\{X \le t\}$ and $F_Y(t) = P\{Y \le t\}$, $t \in (-\infty, \infty)$, are the distribution functions of X and Y, respectively. In case of X and Y are failure times, it is convenient to rewrite the same property as follows:

$$S_X\left(t\right) \le S_Y\left(t\right) \quad \text{for all} \quad t \in \left(-\infty, \infty\right),$$

where $S_X(t) = P\{X > t\}$ and $S_Y(t) = P\{Y > t\}$, $t \in (-\infty, \infty)$, are the survival functions of X and Y, respectively. The relation $X \le^{st} Y$ determines the partial non-strict order on the set of distributions of random variables. In a similar manner, we say the random variables $X_1, X_2, \ldots, X_d$ are completely stochastically ordered

$$X_1 \le^{st} X_2 \le^{st} \ldots \le^{st} X_d$$

if $X_i \le^{st} X_{i+1}$ for $i = 1, \ldots, d - 1$. By the transitivity property of the stochastic order the random variables are stochastically ordered $X_1 \le^{st} X_2 \le^{st} \ldots \le^{st} X_d$, if $X_i \le^{st} X_j$ for all $1 \le i < j \le d$. We say that random variables $X_1, X_2, \ldots, X_d$ are completely stochastically ordered, if there exists a permutation $(\sigma_1, \sigma_2, \ldots, \sigma_d)$ of indices $(1, 2, \ldots, d)$, such that $X_{\sigma_1} \le^{st} X_{\sigma_2} \le^{st} \ldots \le^{st} X_{\sigma_d}$. If the stochastic orders $X_{\sigma_i} \le^{st} X_{\sigma_j}$ hold for some pairs of σ_i and σ_j only, then we report the incomplete stochastic order.

We use a special nonparametric approach to state stochastic orders of failure times in different groups of observations with high reliability. Since the survival function of failure time is equal to 1 at point zero and is tending to 0 as the argument is tending to infinity, the stochastic order cannot be checked in the nonparametric model. The stochastic ordering condition can be relaxed. We say that the random variable X is stochastically smaller than the random variable Y in the weak sense $\left(X \le_\Delta^{st} Y\right)$ with respect to the set Δ, if

$$S_X\left(t\right) \le S_Y\left(t\right) \quad \text{for all} \quad t \in \Delta, \qquad (2)$$

where Δ is some bounded set of positive real numbers. Similarly, the complete stochastic order $X_1 \le^{st} X_2 \le^{st} \ldots \le^{st} X_d$ holds, if

$$S_1(t) \le S_2(t) \le \cdots \le S_d(t) \quad \text{for all} \quad t \in (-\infty, \infty),$$

whereas the corresponding weak stochastic order $X_1 \le^{st}_{\Delta} X_2 \le^{st}_{\Delta} \cdots \le^{st}_{\Delta} X_d$ with respect to Δ holds, if

$$S_1(t) \le S_2(t) \le \cdots \le S_d(t) \quad \text{for all} \quad t \in \Delta,$$

The weak stochastic order $X_1 \le^{st}_{\Delta} X_2 \le^{st}_{\Delta} \cdots \le^{st}_{\Delta} X_d$ can be obtained from $d-1$ pairwise stochastic orders $X_i \le^{st}_{\Delta} X_{i+1}$, $i = 1, \ldots, d-1$, or, in terms of survival functions,

$$S_{X_i}(t) \le S_{X_{i+1}}(t) \quad \text{for all} \quad t \in \Delta \quad \text{and} \quad i = 1, \ldots, d-1.$$

Let $\Delta = \{t_1, t_2, \ldots, t_k\}$ be a finite set. Then (2) can be rewritten as follows:

$$S_X(t_s) \le S_Y(t_s), \quad s = 1, \ldots, k.$$

It seems natural to choose the checkpoints t_s, which cover each of the supports of X and Y, but the statistical framework is inefficient, if the distribution of X is highly biased with respect to Y in this case. In order to increase the efficiency of the statistical analysis, we choose the checkpoints empirically from the combined distribution for each pair of distributions X_i and X_j, $i \ne j$. Let $(\sigma_1, \sigma_2, \ldots, \sigma_s)$, $s \le d$, be an arrangement of the indices $(1, 2, \ldots, d)$; $\left\{\Delta^{(i,j)}\right\}_{i,j=1}^{d} : \Delta^{(i,j)} = \Delta^{(i,j)} \subset (0, \infty)$ is determined for each pair of distributions of X_i and X_j, $i, j \in \{1, \ldots, d\}$. We say the conditional weak stochastic order $X_{\sigma_1} \le^{st}_{\Delta(\sigma_1,\sigma_2)} \cdots \le^{st}_{\Delta(\sigma_{s-1},\sigma_s)} x_{\sigma_s}$ with respect to $\left\{\Delta^{(i,j)}\right\}_{i,j=1}^{d}$ holds, if

$$S_{\sigma_i}(\vec{t}) \le S_{\sigma_j}(\vec{t}) \quad \text{for all} \quad \vec{t} \in \Delta^{(\sigma_1,\sigma_j)}, \quad 1 \le i < j \le s. \tag{3}$$

In other words, the conditional weak stochastic order $X_{\sigma_1} \le^{st}_{\Delta(\sigma_1,\sigma_2)} \cdots \le^{st}_{\Delta(\sigma_{s-1},\sigma_s)} x_{\sigma_s}$ with respect to $\left\{\Delta^{(i,j)}\right\}_{i,j=1}^{d}$ holds, if $S_{\sigma_i} \le^{st}_{\Delta(\sigma_i,\sigma_j)} S_{\sigma_j}$ for all $1 \le i < j \le s$. Note that the pairwise relation of the conditional stochastic order is not transitive in general case, since the stochastic order in each pair is determined in a different weak sense.

Let the survival data contain d groups of independent right-censored observations with failure times T_i having completely unknown survival functions S_i within i-th group, $i = 1, \ldots, d$. Set $\vec{t}^{(i,j)} = \left(t_1^{(i,j)}, t_2^{(i,j)}, \ldots, t_k^{i,j}\right)$ are checkpoints (that can be data based in general case) for weak pairwise stochastic ordering of distributions T_i and T_j for each pair of groups i and j; $\Delta^{(i,j)} = \left(t_1^{(i,j)}, t_2^{(i,j)}, \ldots, t_k^{i,j}\right)$.

We confirm the pairwise weak stochastic order $T_i \le^{st}_{\Delta(i,j)} T_j$ at the confidence level $1 - \alpha$, if each of the particular left sided confidence intervals of level $1 - \alpha/d$ for all of k contrasts $\psi_1^{(i,j)}, \ldots, \psi_k^{(i,j)}$, where $\psi_s^{(i,j)} = S_i\left(t_s^{(i,j)}\right) - S_j\left(t_s^{(i,j)}\right)$, are located entirely to the left of zero. In other words, we obtain joint confidence intervals for the contrasts by using the Bonferroni method. The particular right sided asymptotic confidence intervals are obtained from the asymptotic normality (1). The conditional stochastic order including more than two groups can be confirmed at some confidence level in a similar manner by using the right sided confidence intervals for the contrasts related to all the pairwise orders that determine the conditional stochastic order with the Bonferroni correction on the total number of contrasts.

We also report the p-value that allows to estimate the true confidence of the statistical conclusion. Note that a pairwise weak stochastic order can be confirmed with some confidence, only if (2) holds for the Kaplan–Meier estimators for each $t \in \Delta$. If the pairwise weak stochastic order for the Kaplan–Meier estimators fail, we report the p-value is equal to 1 and the corresponding confidence is estimated equal to 0. Otherwise, the p-value is determined as the infimum of α, such that the pairwise weak stochastic order holds at the confidence level 1 – α. Since a conditional weak stochastic order of 3 and more distributions is determined by the corresponding weak pairwise stochastic orders, the p-value of the conditional weak stochastic order is given as the maximal of p-values of the pairwise stochastic orders. The pairwise stochastic orders are obtained by using the confidence intervals for the contrasts with the correction to the number of weak pairwise stochastic orders that determines the conditional weak stochastic order. Finally, the estimator for the true confidence of a weak stochastic order is equal to 1 – p-value.

Another application of the contrasts method is the detection of available stochastic orders and the confirmation. Note that the whole range of (non-conditional) weak stochastic orders can be determined by using $d(d-1)/2$ pairwise weak stochastic orders and both the alternative pairwise stochastic orders $X_i \leq_{\Delta}^{st} X_j$ and $X_j \leq_{\Delta}^{st} X_i$ related to the same pair (i, j) can be obtained by using the same k contrasts $\psi_s^{(i,j)}$, $s = 1, \dots, k$. If all the two-sided joint confidence intervals for the contrasts lie entirely to the left of zero, we confirm that $T_i \leq_{\Delta(i,j)}^{st} T_j$, whereas if all they lie entirely to the right of zero, we confirm that $T_j \leq_{\Delta(i,j)}^{st} T_i$ with the same confidence as the joint confidence level of the intervals. Hence, only $\dfrac{kd(d-1)}{2}$ contrasts are required to detect all the pairwise weak stochastic orders for any distributions of $T_1, \dots, T_d$.

The conditional weak stochastic orders of 3 and more distributions can be obtained from pairwise weak stochastic orders. We consider the combination of all pairs for which there are the arrangements $(\sigma_1, \sigma_2, \dots, \sigma_s)$, $s \leq d$ of indices $(1, 2, \dots, d)$, such that (3) holds. For example, based on pairwise weak stochastic orders $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_2)}^{st} T_{\sigma_2}$, $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_3)}^{st} T_{\sigma_3}$, $T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3}$, we construct conditional weak stochastic order $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_2)}^{st} T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3}$. In addition, we confirm the following pairwise weak stochastic orders: $T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_4)}^{st} T_{\sigma_4}$, $T_{\sigma_3} \leq_{\Delta(\sigma_3,\sigma_4)}^{st} T_{\sigma_4}$. Then we obtain the conditional weak stochastic orders $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_2)}^{st} T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3}$ and $T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3} \leq_{\Delta(\sigma_3,\sigma_4)}^{st} T_{\sigma_4}$, but not the conditional weak stochastic order $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_2)}^{st} T_{\sigma_2} \leq_{\Delta(\sigma_2,\sigma_3)}^{st} T_{\sigma_3} \leq_{\Delta(\sigma_3,\sigma_4)}^{st} T_{\sigma_4}$, since we do not confirm the pairwise weak stochastic order $T_{\sigma_1} \leq_{\Delta(\sigma_1,\sigma_4)}^{st} T_{\sigma_4}$.

We consider the whole range of the two-sided joint confidence intervals for all $\dfrac{kd(d-1)}{2}$ the contrasts and confirm all available conclusions on the pairwise weak stochastic orders at the confidence level (1 – α). Since we detect conditional stochastic orders, we are ready to reject all inconsistent pairwise stochastic orders.

Statistical data and planning of statistical analysis

The statistical data contains the results of users' jobs processing at the supercomputer center of Peter the Great St. Petersburg Polytechnic University. For each run initiated by the corresponding user's job, we have the processing time of computational task, the number of cores allocated and the exit code, which allows to determine, whether the user's job was completed successfully. Runs of duration less than 5 seconds were removed. Finally, we use information on 1338565 runs from 01.09.2021 to 31.08.2023. Runs that were not completed or not completed successfully are assumed to be censored.

We analyze distributions of the execution time required to complete successfully user's job, in seconds, and the computer time (spent processor time), in processor seconds (sec.* CPU). All user jobs and corresponding runs were classified to 11 groups by user's area of expertise:

- astrophysics;
- bioinformatics;
- biophysics;
- energetics;
- geophysics;
- IT;
- mechanical engineering;
- mechanics;
- physics;
- radiophysics;
- a special group called geovation [10].

The Kaplan–Meier estimators of the survival functions of the required times and computer times to complete successfully user's job are visualized in Figs. 1 and 2.

First, we test the homogeneity null hypothesis that the distributions of the required times (or computer times) to complete successfully user's job in different groups are all the same, as well as the pairwise homogeneity null hypotheses each pair of groups separately by using Wald's type tests. If the null hypothesis of homogeneity is rejected, we perform advanced statistical analysis using contrasts method for each pair of the groups, for which significant differences in distributions of the required times (or computer times) to complete successfully user's job were found, adjusted to the total number of pairs. The conditional weak stochastic orders of 3 and more distributions are obtained from the confirmed pairwise stochastic orders according to (3).

The checkpoints for pairwise homogeneity testing and further advanced analysis of contrasts are obtained in the following way:

1. We obtain $t_{\max}$ is the largest observed failure time of for each of samples.
2. The group with the smaller value of $t_{\max}$ is assumed to be a baseline group.
3. The checkpoints are defined as 7 octiles (12.5%; 25%; 37.5%; 50%; 62.5%; 75%; 87.5%) of the Kaplan–Meier estimator related to the baseline group and the midpoint between the last octile and $t_{\max}$, totally $k = 8$ of the checkpoints.

The checkpoints are consistent estimates of the corresponding numerical characteristics that depend on the joint distribution of failure and censoring times. Then the asymptotic normality of the Kaplan–Meier estimators at the checkpoints is preserved under the null hypothesis and under a fixed alternative.

We use the significance level $\alpha = 0.05$ (5%) and the joint confidence level $1 - \alpha = 0.95$ for all statistical conclusions.

Results of statistical analysis

Testing the homogeneity null hypotheses displays significant differences in distributions of the required times and computer times to complete successfully user's job both with the p-value not exceeding 10^{-300}, the minimal available value in R. Testing the pairwise homogeneity null hypothesis displays highly significant differences in distributions of the required times and computer times to complete successfully user's job for each pair of times and computer times as well with the maximal p-value $8.3 \cdot 10^{-67}$ for times in astrophysics and mechanics groups.

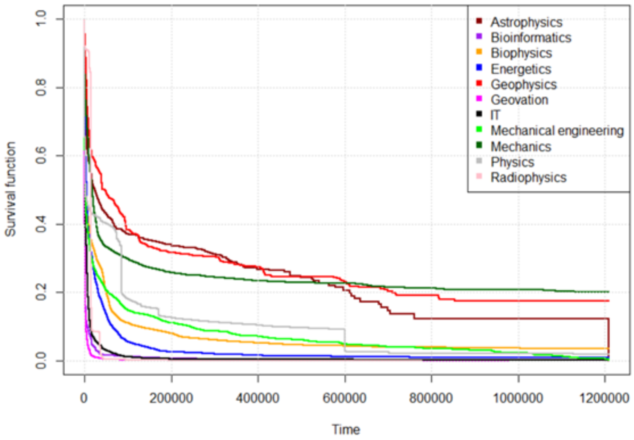


Fig. 1. Kaplan–Meier estimators of the required job execution times

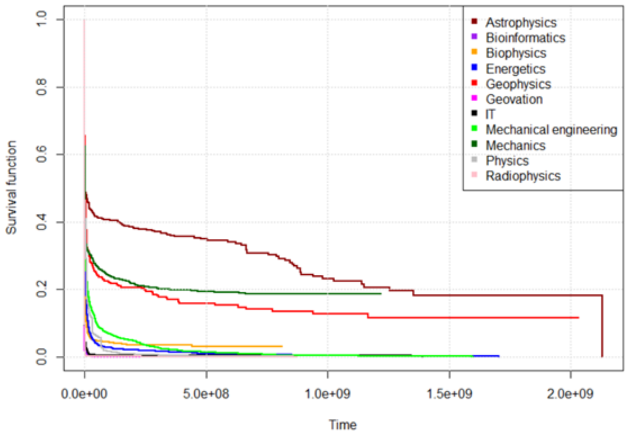


Fig. 2. Kaplan–Meier estimators of the required job execution computer times

The results of detection and confirmation of the obtained pairwise weak stochastic orders in the distributions of the required times and computer times to complete successfully user's job are visualized by using directed graphs in Figs. 3 and 4, respectively: each vertex represents a group by the user's area of expertise; an edge exists, if the stochastic order is confirmed at the joint confidence level of 0.95 adjusted to the total number of pairs; each edge is directed from a larger distribution to a smaller one.

We detect and confirm 21 pairwise weak stochastic orders for the required times to complete successfully user's job and 23 pairwise weak stochastic orders for the required computer times to complete successfully user's job.

The graph shows that we also detect and confirm weak stochastic orders for the three groups, for example, the required times in the geophysics group is stochastically larger than the required times in the biophysics group, which is stochastically larger than that in the geovation group.

We detect and confirm 4 triple weak stochastic orders for the required times to complete successfully user's job and 10 triple weak stochastic orders for the required computer times to complete successfully user's job.

Discussion

In this study, we develop the statistical framework for detection and confirmation, at some confidence level, of weak stochastic orders in distributions of failure times from right-censored survival data.

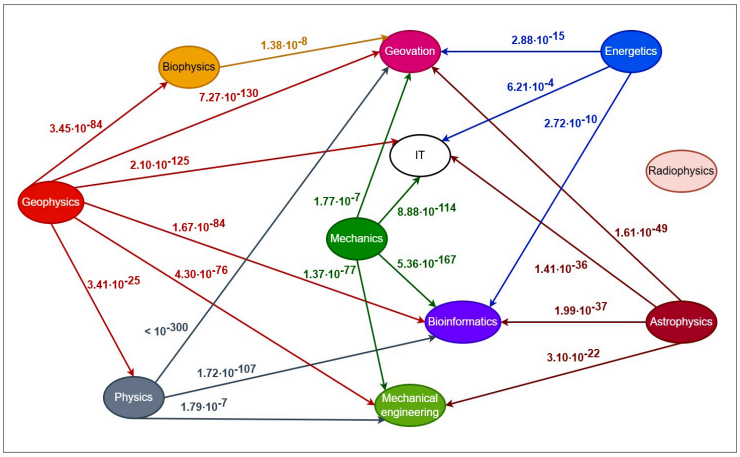


Fig. 3. Significant conditional pairwise weak stochastic orders in distributions of the required times to complete successfully user's job

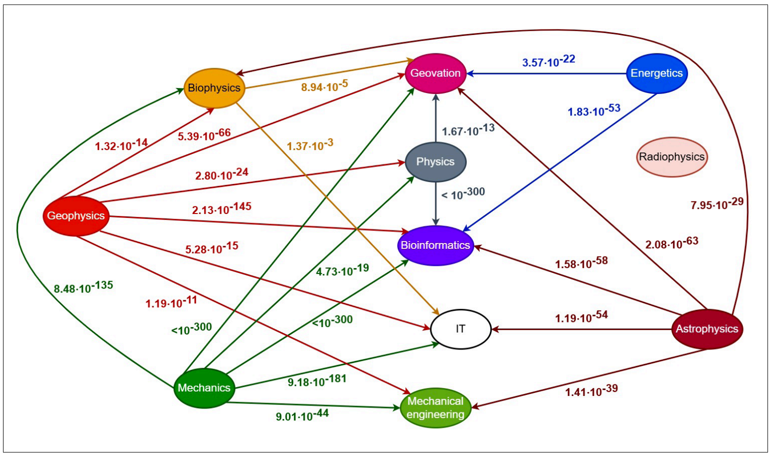


Fig. 4. Significant conditional pairwise weak stochastic orders in distributions of the required computer times to complete successfully user's job

The set of tools for nonparametric categorical analysis of right-censored survival data based on the Kaplan–Meier and the Nelson–Aalen estimators, as well as the contrasts methods for detection and confirmation of weak stochastic orders, were implemented in the R software development environment. The Bonferroni correction is applied to adjust the confidence level for all contrasts under consideration.

We analyze the results of users' jobs processing obtained at the supercomputer center of Peter the Great St. Petersburg Polytechnic University. We group users' jobs into 11 groups by the user's area of expertise. The main objects of interest are the required times and computer times to complete successfully the user's job in different groups of users. We associate these characteristics with the failure time in right-censored data model. Testing the homogeneity null hypotheses of failure time distributions in different groups of users, as well as each of pairwise homogeneity null hypotheses, reveals non-random differences in the corresponding estimators in different groups of users with extremely high significance, close to absolute, for both required times and computer times to complete successfully user's job.

We detect and confirm at the 95% confidence level 21 pairwise weak stochastic orders for the required times to complete successfully user's job and 23 pairwise weak stochastic orders for the required computer times to complete successfully user's job.

Note that the obtained stochastic orders are a much more informative result than simply establishing the significant differences in the distributions. In particular, $T_1 \leq^{st} T_2$ implies that the mean value and all quantiles of T_1 are smaller than the corresponding characteristics of T_2. In some cases, weak stochastic order does not guarantee the existence of a corresponding stochastic order, but it is useful, because it allows us to draw conclusions for the corresponding quantiles. These conclusions are applicable to optimize user's jobs processing.

REFERENCES

1. **Akritas M.G.** Pearson-Type Goodness-of-Fit Tests: The Univariate Case. *Journal of the American Statistical Association*, 1988, Vol. 83, No. 401, Pp. 222–230. DOI: 10.2307/2288944

2. **Bagdonavicius V.B., Nikulin M.S.** Chi-squared goodness-of-fit test for right-censored data. *International Journal of Applied Mathematics & Statistics*, 2011, Vol. 24, No. SI-11A, Pp. 30–50.

3. **Bagdonavičius V., Levuliene R, Nikulin M.S., Tran Q.X.** On chi-square type tests and their applications in survival analysis and reliability. *Journal of Mathematical Sciences*, 2014, Vol. 199, Pp. 88–99. DOI: 10.1007/s10958-014-1835-x

4. **Baranov A.V., Nikolaev D.S.** Machine learning to predict the supercomputer jobs execution time. *Software & Systems*, 2020, Vol. 33, No. 2, Pp. 218–228. DOI: 10.15827/0236-235X.130.218-228.

5. **Gaussier E., Glesser D., Reis V., Trystram D.** Improving backfilling by using machine learning to predict running times. SC'15: *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2015, Pp. 1–10. DOI: 10.1145/2807591.2807646

6. **Guo J., Nomura A., Barton R., Zhang H., Matsuoka S.** Machine learning predictions for underestimation of job runtime on HPC system. In: *Supercomputing Frontiers* (eds. R. Yokota, W. Wu), 2018, Vol .10776. DOI: 10.1007/978-3-319-69953-0_11

7. **Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S.** Random survival forests. *The Annals of Applied Statistics*, 2008, Vol. 2, No. 3, Pp. 841–860. DOI: 10.1214/08-AOAS169

8. **Habib M.G., Thomas D.R.** Chi-square goodness-if-fit tests for randomly censored data. *The Annals of Statistics*, 1986, Vol. 14, No. 2, Pp. 759–765. DOI: 10.1214/aos/1176349953

9. **Hjort N.L.** Goodness of fit tests in models for life history data based on cumulative hazard rates. *The Annals of Statistics*, 1990, Vol. 18, No. 3, Pp. 1221–1258. DOI: 10.1214/aos/1176347748

10. **Hollander M., Peña E.A.** A chi-squared goodness-of-fit test for randomly censored data. *Journal of the American Statistical Association*, 1992, Vol. 87, No. 418, Pp. 458–463. DOI: 10.2307/2290277

11. **Hothorn T, Bühlmann P, Dudoit S, Molinaro A., Van der Laan M.J.** Survival ensembles. *Biostatistics*, 2006, Vol. 7, No. 3, Pp. 355–373. DOI: 10.1093/biostatistics/kxj011

12. **Kim J.H.** Chi-square goodness-of-fit tests for randomly censored data. *The Annals of Statistics*, 1993, Vol. 21, No. 3, Pp. 1621–1639. DOI: 10.1214/aos/1176349275

13. **Kirpichenko S., Utkin L., Konstantinov A., Muliukha V.** BENK: The Beran estimator with neural kernels for estimating the heterogeneous treatment effect. *Algorithms*, 2024, Vol. 17, No. 1, Art. no. 40. DOI: 10.3390/a17010040

14. **Konstantinov A.V.** Predictive models and dynamics of estimates of applied tasks characteristics using machine learning methods. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 3, Pp. 54–60. DOI: 10.18721/JCSTCS.17305

15. **Malov S., O'Brien S.** On survival categorical methods with applications in epidemiology and AIDS research. In: *Applied Methods of Statistical Analysis. Applications in Survival Analysis, Reliability and Quality Control* (eds. B. Lemeshko, M. Nikulin, N. Balakrishnan), 2013, Pp. 173–180.

16. **Malov S.V., Lukashin A.A.** Count time series analysis of jobs scheduling in the hybrid supercomputer center. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 3, Pp. 42–53. DOI: 10.18721/JCSTCS.17304

17. **McKenna R., Herbein S., Moody A., Gamblin T., Taufer M.** Machine learning predictions of runtime and IO traffic on high-end clusters. *2016 IEEE International Conference on Cluster Computing* (*CLUSTER*), 2016, Pp. 255–258. DOI: 10.1109/CLUSTER.2016.58

18. **Savin G.I., Shabanov B.M., Nikolaev D.S., Baranov A.V., Telegin P.N.** Jobs runtime forecast for JSCC RAS supercomputers using machine learning methods. *Lobachevskii Journal of Mathematics*, 2020, Vol. 41, Pp. 2593–2602. DOI: 10.1134/S1995080220120343

19. **Hu S., Fridgeirsson E., van Wingen G., Welling M.** Transformer-based deep survival analysis. *Proceedings of AAAI Spring Symposium on Survival Prediction – Algorithms, Challenges and Applications*, 2021, Vol. PLMR 146, Pp. 132–148.

20. **Turnbull B.W., Weiss L.** A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, 1978, Vol. 34, No. 3, Pp. 367–375.

21. **Utkin L.V., Konstantinov A.V., Eremenko D.Yu. et al.** Interpretation methods for machine learning models in the framework of survival analysis with censored data: a brief overview. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 3, Pp. 22–31. DOI: 10.18721/JCSTCS.17302

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Misharina Tatiana A.
Мишарина Татьяна Андреевна
E-mail: tanechkamisharina254@gmail.com

Malov Sergey V.
Малов Сергей Васильевич
E-mail: sergey.v.malov@gmail.com
ORCID: https://orcid.org/0000-0003-0093-6506

Submitted: 09.11.2024; Approved: 21.04.2025; Accepted: 13.05.2025.
Поступила: 09.11.2024; Одобрена: 21.04.2025; Принята: 13.05.2025.

Research article
DOI: https://doi.org/10.18721/JCSTCS.18202
UDC 622.001

SYSTEM OF INTERCONNECTED SOLUTIONS "INTELLIGENT QUARRY"

V.Yu. Bazhin[1], *A.S. Anufriev*[2] ✉ , *L.A. Rusinov*[3]

[1] Empress Catherine II Saint Petersburg Mining University,
St. Petersburg, Russian Federation;

[2] Engineering Laboratory LLC, St. Petersburg, Russian Federation;

[3] Saint-Petersburg State Institute of Technology (SPSIT),
St. Petersburg, Russian Federation

✉ anufriev_rf@yahoo.com

Abstract. The article discusses the implementation of digital solutions in the management system of the mining and transportation complex using the case of Karelsky Okatysh JSC. An analysis of the initial state of the enterprise's technological chain is presented, highlighting key issues related to the stability of the blend composition, the quality of the mined ore and the efficiency of its transportation management. To address these problems, the software and hardware system called "Intelligent Quarry" was developed, comprising interconnected modules for blend stabilization, automated raw material quality monitoring and predictive equipment condition control. The effectiveness of the proposed solutions was confirmed by simulation of the expected results and subsequent comparison of production indicators before and after the system's implementation. A technical and economic analysis confirmed the increase in concentrate output by 0.84%, a decrease in magnetic iron content in tailings by 0.2% and an additional annual concentrate yield of 171438 tons.

Keywords: mining and transportation complex, intelligent mine, mining and processing plant, automated management systems, hyperspectral sensing, neural networks

Citation: Bazhin V.Yu., Anufriev A.S., Rusinov L.A. System of interconnected solutions "Intelligent Quarry". Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 21–32. DOI: 10.18721/JCSTCS.18202

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18202
УДК 622.001

СИСТЕМА ВЗАИМОСВЯЗАННЫХ РЕШЕНИЙ «РАЗУМНЫЙ КАРЬЕР»

В.Ю. Бажин[1], *А.С. Ануфриев*[2] ✉ , *Л.А. Русинов*[3]

[1] Санкт-Петербургский горный университет имени императрицы Екатерины II, Санкт-Петербург, Российская Федерация;

[2] ООО «Лаборатория Инжиниринга», Санкт-Петербург, Российская Федерация;

[3] Санкт-Петербургский государственный технологический институт (технический университет), Санкт-Петербург, Российская Федерация

✉ anufriev_rf@yahoo.com

Аннотация. В статье рассмотрены вопросы внедренных цифровых решений в систему управления горнотранспортным комплексом на примере АО «Карельский окатыш». Представлен анализ исходного состояния технологической цепочки предприятия, выявлены основные проблемы, связанные со стабильностью состава шихты, качеством добываемой руды и эффективностью управления ее транспортировкой. Для решения указанных проблем был разработан программно-технический комплекс «Разумный карьер», включающий взаимосвязанные модули стабилизации шихты, автоматизированного мониторинга качества сырья и предиктивного контроля состояния оборудования. Эффективность предложенных решений подтверждена с помощью моделирования ожидаемых эффектов и последующего сравнения производственных показателей до и после внедрения комплекса. Технико-экономический анализ подтвердил увеличение производительности по концентрату на 0,84%, снижение содержания магнитного железа в хвостах на 0,2% и дополнительный годовой выход концентрата в объеме 171438 тонн.

Ключевые слова: горнотранспортный комплекс, интеллектуальный рудник, горно-обогатительный комбинат, автоматизированные системы управления, гиперспектральное зондирование, нейронные сети

Для цитирования: Bazhin V.Yu., Anufriev A.S., Rusinov L.A. System of interconnected solutions “Intelligent Quarry” // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 21–32. DOI: 10.18721/JCSTCS.18202

Introduction

The energy costs of open-pit mining enterprises are significant, with mining technological processes accounting for approximately half of these expenses. Quarry excavators, which perform more than 80% of the total volume of operations, consume a considerable amount of electricity. Therefore, the quality of their performance largely determines the overall efficiency of the mining enterprise.

The modernization of the control system for the mining and transportation complex (MTC) is a key stage in the development of the modern mining industry. In recent years, there has been significant progress in the field of automation and digitalization of mining enterprises. The evolution of technologies, from advanced data analytics to artificial intelligence (AI), has enormous potential to transform the mining industry by improving operational efficiency, productivity and production safety.

The totality of methods for automation and optimization of all aspects of mineral extraction is reflected in the concept of the *intelligent quarry*. The main components of intelligent mines include equipment automation, real-time monitoring systems, data analytics, AI, digital twins and advanced safety management systems [1–3].

Review of existing solutions in the field of integrated automation of ore mining and transportation processes

Below are the examples of successful implementations of integrated automation in various regions of the world.

1. *Rio Tinto's "Mine of the Future" program*

The Australian company Rio Tinto implemented the Mine of the Future™ program aimed at creating fully automated mines. As part of this project, autonomous haul trucks, automated drilling rigs, and AutoHaul™ system, the world's first fully autonomous railway network for iron ore transportation, were deployed at mines in the Pilbara region. These solutions led to a 15% increase in productivity and a 13% reduction in operating costs [4].

2. *Intelligent control systems at BHP mines*

BHP has implemented intelligent dispatch systems at the Jimblebar and South Flank mines in Australia. These systems use geoinformation technologies and predictive analytics algorithms to optimize the routing of haul trucks, which has resulted in a 25% reduction in equipment downtime and increased accuracy in mine planning [5].

3. *Digital transformation of Vale mines*

The Brazilian company Vale is actively implementing the Internet of things (IoT) technologies, digital twins and hyperspectral sensors for real-time monitoring of ore quality. This improves the accuracy of forecasting the composition of extracted raw materials and reduces losses during processing. In addition, the use of predictive analytics has contributed to a 30% reduction in equipment maintenance costs[1].

4. *Kankberg smart mine by Boliden*

The Swedish company Boliden implemented a smart mine project at the Kankberg site, where a 5G network was deployed for wireless control of underground equipment. The use of digital twins and automated monitoring systems resulted in a 10% increase in productivity and a 15% reduction in operating costs[2].

5. *Norilsk Nickel's "Smart Mine" project*

The Russian company Norilsk Nickel is developing an autonomous "unmanned" mine project, which employs computer vision and AI technologies for monitoring and controlling mining processes. One of the mines of the Polar Division was selected as a test site for the implementation of autonomous systems[3].

6. *Application of AI in ore beneficiation at KAZ Minerals enterprises*

KAZ Minerals has introduced an AI-based tool to optimize the ore beneficiation process. The system analyzes data from the mine face to the tailings storage facility, using more than 500 million experimental data points, and is capable of self-learning, offering recommendations to improve technological processes[4].

Despite significant achievements in the automation of ore mining and transportation, the implementation of digital technologies is accompanied by a number of technological, economic and organizational limitations. An analysis of existing projects shows that the comprehensive integration of digital solutions in the mining sector remains a complex task requiring a systematic approach.

One of the key limitations is the high capital intensity of digitalization. The implementation of autonomous systems, digital twins and predictive analytics requires substantial investments in equipment

[1] Vale launches innovative program for digital transformation of its supply chain, Available https://vale.com/w/vale-launches-innovative-program-for-digital-transformation-of-its-supply-chain (Accessed 10.03.2025)

[2] Voigt B., Falshaw S. Boliden Summary Report, Resources and Reserves 2024, Kankberg, Available: https://www.boliden.com/490349/globalassets/operations/exploration/mineral-resources-and-mineral-reserves-pdf/2024/resources-and-reserves-kankberg-2024-12-31.pdf (Accessed 13.05.2025)

[3] Innovatsii i tsifrovye tekhnologii [Innovation and digital technologies], Available: https://ar2023.nornickel.ru/business-overview/innovation-digital-technologies (Accessed 10.03.2025)

[4] Iskusstvennyi intellekt [Artificial intelligent], Available: https://www.kazminerals.com/ru/repository/news-container/news/2021/искусственный-интеллект/ (Accessed 10.03.2025)

modernization, sensor systems and data collection infrastructure. In the context of commodity market volatility, enterprises are forced to limit the scale of digitalization, focusing on specific nodes of the technological chain.

An additional problem is the difficulty of integrating digital solutions with outdated infrastructure. Most existing mining complexes use equipment installed several decades ago, which complicates the adaptation of modern automated systems. The integration of autonomous haul trucks and drilling rigs into existing dispatch systems requires significant modifications, increasing the overall cost of digitalization.

Furthermore, automation remains fragmented. Most projects are focused on individual aspects, such as transportation operations, intelligent data analytics, or predictive maintenance of equipment [6–8].

To minimize the identified drawbacks of the existing approaches to the automation of mining and transportation processes, the authors propose the concept of an "Intelligent Quarry." This system is an adaptive platform that includes automated monitoring systems, autonomous haul equipment, intelligent control algorithms and predictive analytics.

A key feature of the concept is the modularity of its architecture, which enables the phased implementation of individual components depending on the production conditions of a particular enterprise. This allows digital solutions to be integrated without a complete infrastructure overhaul, adapting the system to the changing requirements of the technological process. The flexibility and scalability of the approach ensure its applicability at enterprises with varying levels of automation, reducing implementation costs and increasing the efficiency of digital transformation in the mining industry [9].

Audit of the MTC of the mining and processing plant

For conducting industrial trials and subsequent implementation, *Karelsky Okatysh* JSC was selected – one of the largest mining enterprises in the north-west of Russia with significant development potential. The company's core activity is the extraction and processing of ferruginous quartzites into high-quality iron ore raw materials, namely pellets [10].

The enterprise's MTC includes several main components: ore extraction, transportation, processing and storage. To identify bottlenecks in the technological chain and to form the most effective system of interconnected solutions, an audit of the MTC was conducted, the results of which are presented in Table 1.

The enterprise's main request was the modernization of the MTC in order to increase the volume of iron ore concentrate depending on the properties of the ore in the flow. This includes identifying methods for increasing concentrate output and reducing production losses by improving the accuracy of information on the incoming ore. Refining ore characteristics, such as iron and sulfur content and beneficiation potential, from the moment of extraction to delivery at the crushing and beneficiation plant (CBP) is also a critical task.

To address the key issues identified during the audit, the proposed system is presented as a set of interconnected solutions encompassing the entire MTC production process. This approach ensures integrity and continuity throughout the production chain, while the modular structure provides flexibility and adaptability to the specific conditions and technologies of each enterprise [11–13].

Integrated set of interconnected solutions

The comprehensive system of interconnected solutions for quality planning and ore blending covers the entire production process of ore extraction, transportation and storage. This approach ensures the integrity and continuity of the production chain. The modular structure of the hardware and software system (HSS) provides flexibility and adaptability to the conditions and technologies specific to each enterprise.

To achieve the main goal of implementing the system, namely, increasing iron extraction regardless of the type of incoming ore, the HSS must aggregate information at all levels and manage processes either in automatic mode or through recommendation-based control.

The solutions developed within the "*Intelligent Quarry*" HSS can be divided into two groups depending on the performed task: stabilization or optimization of the ore blend (Table 2).

Table 1

Bottlenecks in the MTC of Karelsky Okatysh JSC

Main processes	Identified bottlenecks
Ore extraction	– Insufficient accuracy of data in the block model of the rock mass. – Lack of automatic solutions for the analysis of the surface layer of the mining face. – Inaccuracies in excavation and extraction processes.
Ore transportation	– Ore samples at the ore control station (OCS) are not always representative. – Insufficient automation of haul truck movement in the cyclic-flow technology (CFT) area. – Problems with predicting the ore loading time into dump cars.
Ore processing	– Insufficient reliability of information on the qualitative characteristics of ore at transfer stockpiles. – Deficiencies in forming the gradient of ore quality indicators. – Problems with accounting for oversized material.
Storage and logistics	– High time expenditures at the planning stage and when forming the delivery order to the plant. – Poor optimization of in-pit logistics. – Difficulties in forecasting outcomes under emergencies. – Inaccuracies and incompleteness in the database.
Ore blending	– Existing ore blending algorithms are suboptimal and require refinement to account for ore beneficiation potential. – Current blending methods do not adequately consider ore processability.
Information systems and accounting	– The enterprise database contains inaccuracies and incomplete information. – The use of different data calculation algorithms across departments results in contradictions. – Deficiencies in methods for constructing the blasted ore model from the block model.

Figs. 1 and 2 present diagrams illustrating the key components of the "Intelligent Quarry" HSS and their interconnections with technological processes, from geological exploration to railway transportation.

– **Geological Exploration → Drilling and Blasting → Excavation**

At the initial stages of mining, the *Block-to-Blast Model Conversion* module is used to transform the block model into a blast model, which is applied in both long-term and operational planning [14, 15]. This information is then transferred to the drilling and blasting planning stage (Fig. 1, *a*).

– **Excavation**

After blasting, the *In-Pit Transportation Accounting* module tracks the delivery of backfill material for road and excavator pad preparation, and records transportation data and ore conditions at the block in case of emergency. The same *Block-to-Blast Model Conversion* module is used for planning excavator operations, forming a digging map based on its results. During excavation, the *Ore Quality Recognition* module using a Fourier interferometer is applied to distinguish overburden from ore and collect data on ore quality characteristics [16–18]. This data is transferred to the *Stockpile Formation Algorithm* and *Blend Formation* modules to generate dispatch plans for ore delivery to the processing plant (Fig. 1, *b*).

– **Ore Transportation by Haul Trucks**

After ore is loaded, the haul truck moves to the Ore Control Station (OCS), where the *Ore Mass Measurement in Haul Trucks* module estimates the ore mass based on volumetric fill level, increasing accuracy in quantitative ore accounting. In the case of an emergency dump within the pit, the *In-Pit Transportation Accounting* module records ore presence on the block and sends this information to the duty geologist for further decision making. Upon unloading at the transfer stockpile, the *Stockpile Formation Algorithm* calculates the quality gradient based on the unloading location, refining ore characteristics for the *Railcar Movement Scheduling* module, which plans rail dispatch from the transfer stockpile [19–21] (Fig. 1, *c*).

Table 2

Description of the modules of the "Intelligent Quarry" HSS

Module	Tasks	Description
Stabilization of the blend		
Conversion of block model to blast model	– Reduce specific consumption of explosives through adaptive drilling and blasting grid planning based on rock characteristics	This module uses neural networks to analyze geological data and drilling parameters. Inputs include physical and mechanical properties of rocks, explosive parameters and drilling configurations
Fourier spectrometer	– Speed up ore quality assessment by visualizing in-blast ore characteristics. – Improve ore quality data in haul trucks	The module uses interferometer-based Fourier spectrometry for remote sensing and in-pit ore visualization. It allows distinguishing ore from waste and collecting data on valuable components and impurities
In-pit transportation accounting	– Improve planning quality of new block development. – Reduce transport costs through accurate haul truck load tracking	Uses GNSS for real-time vehicle tracking and RFID for automatic vehicle identification. The module accounts for vehicle location and load status in different quarry zones
Ore mass measurement in haul trucks	– Improve mass accounting for ore transported to stockpiles. – Improve ore inventory accuracy at stockpiles	Uses LiDAR to measure distances via laser beam. LiDAR scanners installed at the OCS capture 3D profiles of truck beds to determine transported ore volume and mass
Stockpile formation algorithm	– Improve ore inventory accuracy at stockpiles. – Improve information on ore properties dispatched to the plant	Based on data integration: ore properties, logistics, and production plans. It accounts for stockpile dynamics and ore distribution strategies
Railcar scheduling module	– Improve planning quality by accounting for organizational factors. – Automate the railcar dispatching process	Operates on real-time data about ore availability, stockpile status, and plant operations. It generates production plans and optimized transport schedules
Granulometric composition control	– Improve drilling and blasting planning via analysis of grid structure and actual particle size data	Analyzes drilling and blasting grid and granulometric composition using machine vision. This improves drilling patterns and explosive usage based on rock fragmentation
Stockpile positioning system	– Improve accuracy of ore property information in formed stockpiles	Uses radio frequency sensors to refine ore quality tracking in formed stockpiles
Optimization of the blend		
Blend formation module	– Reduce losses via optimal processing modes for specific ore types. – Reduce risks of losing magnetic properties due to overgrinding	Uses AI-based modeling to simulate ore processing, adjusting production parameters. Targets maximum output under given constraints by combining data on ore quality, stockpile state, and transport logistics
Calculator	– Improve ore mass accounting for material transported to stockpiles	Using a mathematical modeling apparatus, the module allows monitoring production processes in real time, without being tied to three-hour cycles of chemical analysis, but using this information as calibration.

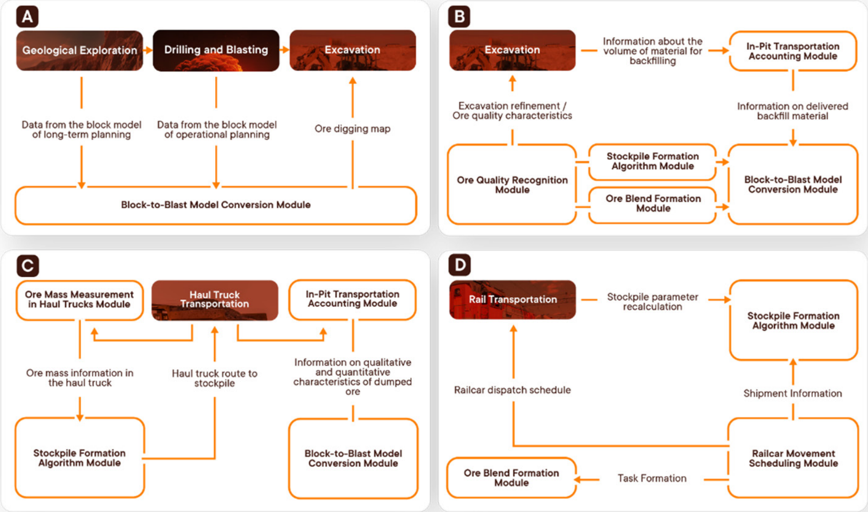

A
Geological Exploration
Drilling and Blasting
Excavation
Data from the block model of long-term planning
Data from the block model of operational planning
Ore digging map
Block-to-Blast Model Conversion Module
B
Excavation
Information about the volume of material for backfilling
In-Pit Transportation Accounting Module
Excavation refinement / Ore quality characteristics
Information on delivered backfill material
Ore Quality Recognition Module
Stockpile Formation Algorithm Module
Ore Blend Formation Module
Block-to-Blast Model Conversion Module
C
Ore Mass Measurement in Haul Trucks Module
Haul Truck Transportation
In-Pit Transportation Accounting Module
Ore mass information in the haul truck
Haul truck route to stockpile
Information on qualitative and quantitative characteristics of dumped ore
Stockpile Formation Algorithm Module
Block-to-Blast Model Conversion Module
D
Rail Transportation
Stockpile parameter recalculation
Stockpile Formation Algorithm Module
Railcar dispatch schedule
Shipment Information
Railcar Movement Scheduling Module
Ore Blend Formation Module
Task Formation

Fig. 1. Interconnection of modules in the “Intelligent Quarry” HSS

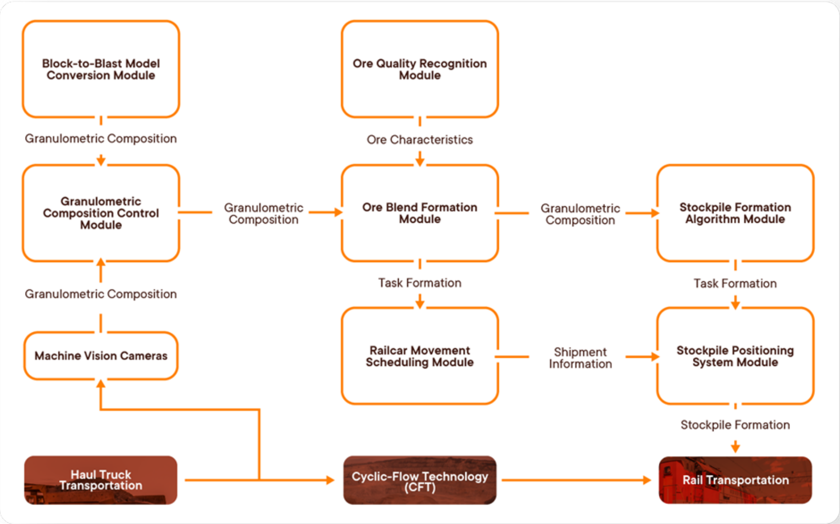

Block-to-Blast Model Conversion Module
Ore Quality Recognition Module
Granulometric Composition
Ore Characteristics
Granulometric Composition Control Module
Granulometric Composition
Ore Blend Formation Module
Granulometric Composition
Stockpile Formation Algorithm Module
Granulometric Composition
Task Formation
Task Formation
Machine Vision Cameras
Railcar Movement Scheduling Module
Shipment Information
Stockpile Positioning System Module
Stockpile Formation
Haul Truck Transportation
Cyclic-Flow Technology (CFT)
Rail Transportation

Fig. 2. Interconnection of modules in the CFT framework

– **Ore Transportation by Railcars**

During scheduling of railcar dispatch to transfer stockpiles, the *Railcar Movement Scheduling* module calculates the number of trains, their travel, loading, and unloading times, taking into account organizational constraints and data from the *Blend Formation* module, which forms ore “packages” considering iron and sulfur content for further crushing [22].

Ore shipment from the stockpile is accompanied by recalculation of ore characteristics in the *Stockpile Formation Algorithm* module (Fig. 1, *d*).

– **Cyclic-Flow Technology (CFT)**

When operating at the central quarry, CFT will be used for the transportation of ore and overburden. To monitor ore quality characteristics, the *Granulometric Composition Control* module collects grain size data from the *Block-to-Blast Model Conversion* module and technical vision cameras [23–25].

The *Ore Quality Recognition* module or sensors of the ore stream monitoring system provide real-time control of ore quality parameters.

The *Stockpile Formation Algorithm* module forms ore piles based on data from the *Blend Formation* module. The positioning of the stacking machine is carried out using a radio-frequency-based positioning system [26] (Fig. 2).

Results and discussion

To justify the feasibility of implementing the proposed solutions, a techno-economic assessment (TEA) was conducted. This included evaluating the mutual influence of solutions on one another. A methodology was proposed that combines expert evaluation, statistical methods and combinatorics. Table 3 presents the list of solutions along with their weighting coefficients [27, 28].

Table 3

Weighting coefficients of solutions in the "Intelligent Quarry"

№	Solution	Weighting coefficient
1. *Blend Optimization*		
1.1	Blend Formation Module	0.80
1.2	Calculator Module	0.20
2. *Blend Stabilization*		
2.1	Block-to-Blast Model Conversion Module	0.10
2.2	Imaging Fourier Spectrometer	0.30
2.3	In-Pit Transportation Accounting Module	0.10
2.4	Ore Mass Measurement in Haul Trucks	0.20
2.5	Stockpile Formation Algorithm Module	0.15
2.6	Railcar Scheduling Module	0.05
2.7	Granulometric Composition Control Module	0.05
2.8	Stockpile Positioning System	0.05

The normalized effect for each group of solutions is determined by the following equation:

$$Э_i = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix} \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} Q_i, \tag{1}$$

where $Э_i$ is the normalized effect for the i-th group of solutions; γ_n is the weighting coefficient of the solution; Q_i is the production increase in concentrate for the i-th group.

If a solution is selected from a group, a "1" is assigned in the identity matrix row corresponding to the relevant weighting coefficient.

The total normalized effect, considering the mutual influence of the solutions, is calculated by the equation:

$$Э_\Sigma = \sum_{i=1}^{3} Э_i + \Delta Q \cdot \sum_{i=1}^{3} \alpha_i \frac{n_i}{N_i}, \tag{2}$$

where ΔQ is the production increase from the synergistic effect; α_i is the influence degree of the weighting coefficient (see Table 4); n_i is the number of solutions implemented from the group; N_i is the total number of solutions in the group [29, 30].

Table 4

Influence degree of weighting coefficients

Influence Degree	Value	Condition
α_1	0.1	if $\gamma_n \in [0; 0.1]$
α_2	0.3	if $\gamma_n \in [0.1; 0.2]$
α_3	0.6	if $\gamma_n \in [0.2; 1.0]$

The increase in concentrate production from the synergistic effect is calculated as:

$$\Delta Q = Q_{\Sigma} - \sum_{i=1}^{3} Q_i, \tag{3}$$

where Q_{Σ} is the increase in concentrate production.

Thus, using the formula for the total normalized effect, it is possible to calculate the total technical effect depending on the decisions taken for implementation.

To conduct a quantitative assessment, key production indicators were selected, such as ore and concentrate productivity, magnetic iron content in tailings, product yield and extraction [31, 32].

The implementation of the "Intelligent Quarry" HSS is expected to result in: an increase in productivity from 356 to 359 t/h, an increase in iron extraction by 0.61% and a decrease in the magnetic iron content in tailings from 1.72% to 1.52%, which indicates an improvement in the beneficiation efficiency.

The expected increase in the yield of the final product was 0.2%, and the annual concentrate production was 171438 tons, which confirms an increase in technological efficiency due to the optimization of ore flow management processes and stabilization of the blend parameters [33, 34].

Conclusion

The concept of the "Intelligent Quarry" HSS demonstrates high potential for implementation across various mining and processing plants due to its adaptability to production conditions and modular architecture. This enables enterprises to gradually integrate digital technologies without the need for complete infrastructure modernization, minimizing costs and reducing associated risks.

The system's flexibility allows for customization to match specific parameters of ore extraction, transportation, and processing, making it applicable to large-scale mining and processing plants as well as medium and small enterprises. Further development of the system may include the expansion of predictive analytics, digital orebody modeling and automated equipment control, providing opportunities to optimize production processes, increase recovery rates of valuable components and reduce technological losses.

The prospects for deploying the *Intelligent Quarry* system are associated with the continued expansion of its functionality, including the advanced use of predictive analytics, integration with digital twins of deposits and the implementation of autonomous equipment control systems. The evolution of AI and the IoT technologies creates the groundwork for establishing a unified intelligent environment that unites mining, transportation and processing into a closed-loop automated circuit. This will not only improve operational efficiency, but also reduce environmental impact through the rational use of resources and optimization of energy consumption.

REFERENCES

1. **Noriega R., Pourrahimian Y.** A systematic review of artificial intelligence and data-driven approaches in strategic open-pit mine planning. *Resources Policy*, 2022, Vol. 77, Art. no. 102727. DOI: 10.1016/j.resourpol.2022.102727

2. **Molaei F., Rahimi E., Siavoshi H., Afrouz S.G., Tenorio V.** A comprehensive review on internet of things (IoT) and its implications in the mining industry. *American Journal of Engineering and Applied Sciences*, 2020, Vol. 13, No. 3, Pp. 499–515. DOI: 10.3844/ajeassp.2020.499.515

3. **Elbazi N., Mabrouki M., Chebak A., Hammouch F.** Digital twin architecture for mining industry: Case study of a stacker machine in an experimental open-pit mine. *2022 4th Global Power, Energy and Communication Conference* (*GPECOM*), 2022, Pp. 232–237. DOI: 10.1109/GPECOM55404.2022.9815618

4. **Marenge F., Montgomery J., Zaw M., Marey Y.Y.** Case Study: Rio Tinto & Mine of the Future. LB5230 Managing Strategic Resources and Operations. James Cook University, 2018. 21 p., Available: https://www.academia.edu/29037773/Rio_Tinto_and_Mine_of_the_Future_LB5230_Managing_Strategic_Resources_and_Operations (Accessed 13.05.2025)

5. **Cehlar M., Zhironkin S.A., Zhironkina O.V.** Digital technologies of industry 4.0 in mining 4.0 – prospects for the development of geotechnology in the XXI century. *Bulletin of the Kuzbass State Technical University*, 2020, Vol. 139, No. 3, Pp. 80–90. DOI: 10.26730/1999-4125-2020-3-80-90

6. **Abdellah W.R., Kim J.-G., Hassan M.M.A., Ali M.A.M.** The key challenges towards the effective implementation of digital transformation in the mining industry. *Geosystem Engineering*, 2022, Vol. 25, No. 1–2, Pp. 44–52. DOI: 10.1080/12269328.2022.2120093

7. **Sánchez F., Hartlieb P.** Innovation in the mining industry: Technological trends and a case study of the challenges of disruptive innovation. *Mining, Metallurgy & Exploration*, 2020, Vol. 37, Pp. 1385–1399. DOI: 10.1007/s42461-020-00262-1

8. **Zhukovskiy Y.L., Semenyuk A.V., Alieva L.Z., Arapova E.G.** Blockchain-based digital platforms to reduce the carbon footprint of mining. *Mining informational and analytical bulletin* (*MIAB*), 2022, Vol. 6–1, Pp. 361–378. DOI: 10.25018/0236_1493_2022_61_0_361

9. **Koteleva N., Khokhlov S., Frenkel I.** Digitalization in open-pit mining: A new approach in monitoring and control of rock fragmentation. *Applied Sciences*, 2021, Vol. 22, No. 11, Art. no. 10848. DOI: 10.3390/app112210848

10. **Martins P., Soofastaei A.** Predictive maintenance of mining machines applying advanced data analysis. In: *Data Analytics Applied to the Mining Industry*. Boca Raton: CRC Press, 2020. Pp. 149–168. DOI: 10.1201/9780429433368

11. **Dayo-Olupona O., Genc B., Celik T., Bada S.** Adoptable approaches to predictive maintenance in mining industry: An overview. *Resources Policy*, 2023, Vol. 86, Part A, Art. no. 104291. DOI: 10.1016/j.resourpol.2023.104291

12. **Rylnikova M.V., Klebanov D.A., Makeev M.A., Kadochnikov M.V.** Application of artificial intelligence and the future of big data analytics in the mining industry. *Russian Mining Industry*, 2022, Vol. 3, Pp. 89–92. DOI: 10.30686/1609-9192-2022-3-89-92

13. **Myshletsov A.I., Avrutskaya S.G.** Introduction of digital technologies in mining industry. *Uspekhi v Khimii i Khimicheskoi Tekhnologii Magazine*, 2022, Vol. 36, No. 1, Pp. 70–73.

14. **Lukichev S.V.** Digital past, present, and future of mining industry. *Russian Mining Industry*, 2021, Vol. 4, Pp. 73–79. DOI: 10.30686/1609-9192-2021-4-73-79

15. **Li L., Li Y., Zhang X., He Y., Yang J., Tian B., Ai Y., Li L., Nüchter A., Xuanyuan Z.** Embodied intelligence in mining: Leveraging multi-modal large language model for autonomous driving in mines. *IEEE Transactions on Intelligent Vehicles*, 2024, Vol. 9, No. 5, Pp. 4831–4834. DOI: 10.1109/TIV.2024.3417938

16. **Hazrathosseini A., Moradi Afrapoli A.** Maximizing mining operations: Unlocking the crucial role of intelligent fleet management systems in surface mining's value chain. *Mining*, 2024, Vol. 4, No. 1, Pp. 7–20. DOI: 10.3390/mining4010002

17. **Zhang K., Kang L., Chen X., He M., Zhu C., Li D.** A review of intelligent unmanned mining current situation and development trend. *Energies*, 2022, Vol. 15, No. 2, Art. no. 513. DOI: 10.3390/en15020513

18. **Mialeshka Yu.V.** Digitalization of the mining enterprise as a factor of ensuring its economic security. *Technical and Technological Problems of the Service*, 2020, Vol. 52, No. 2, Pp. 59–63.

19. **Zimovets A.V., Klimachev T.D.** Digital transformation of production at Russian enterprises under import substitution policy. *Russian Journal of Innovation Economics*, 2022, Vol. 12, No. 3, Pp. 1409–1426. DOI: 10.18334/vinec.12.3.116297

20. **Matevosian R.A., Varfolomeyev I.A.** Software for the analysis and control model of the sinter charge composition. *Cherepovets State University Bulletin*, 2022, Vol. 111, No. 6, Pp. 65–78. DOI: 10.23859/1994-0637-2022-6-111-5

21. **Pelevin A.E.** Iron ore beneficiation technologies in Russia and ways to improve their efficiency. *Journal of Mining Institute*, 2022, Vol. 256, Pp. 579–592. DOI: 10.31897/PMI.2022.61

22. **Kobzev V.V., Babkin A.V., Skorobogatov A.S.** Digital transformation of industrial enterprises in the new reality. *π-Economy*, 2022, Vol. 15, No. 5, Pp. 7–27. DOI: 10.18721/JE.15501

23. **Krakovskaya I.N.** The concept of sustainable competitiveness of industrial clusters in Russia: the main provisions. *Journal of Economics, Entrepreneurship and Law*, 2023, Vol. 13, No. 2, Pp. 343–364. DOI: 10.18334/epp.13.2.116984

24. **Gospodarikov A.P., Revin I.E., Morozov K.V.** Composite model of seismic monitoring data analysis during mining operations on the example of the Kukisvumchorrskoye deposit of AO Apatit. *Journal of Mining Institute*, 2023, Vol. 262, Pp. 571–580. DOI: 10.31897/PMI.2023.9

25. **Kulchitskiy A.A., Mansurova O.K., Nikolaev M.Yu.** Recognition of defects in hoisting ropes of metallurgical equipment by an optical method using neural networks. *Chernye Metally*, 2023, Vol. 3, Pp. 81–88. DOI: 10.17580/chm.2023.03.13

26. **Saadoun A., Fredj M., Boukarm R., Hadji R.** Fragmentation analysis using digital image processing and empirical model (KuzRam): a comparative study. *Journal of Mining Institute*, 2022, Vol. 257, Pp. 822–832. DOI: 10.31897/PMI.2022.84

27. **Zakharov V.N., Kubrin S.S.** Digital transformation and intellectualization of mining systems. *Mining Informational and Analytical Bulletin* (*MIAB*), 2022, Vol. 5–2, Pp. 31–47. DOI: 10.25018/0236_1493_2022_52_0_31

28. **Nevskaya M.A., Raikhlin S.M., Chanysheva A.F.** Assessment of energy efficiency projects at russian mining enterprises within the framework of sustainable development. *Sustainability*, 2024, Vol. 16, No. 17, Art. no. 7478. DOI: 10.3390/su16177478

29. **Aleksandrova T., Nikolaeva N., Afanasova A., Romashev A., Kuznetsov V.** Justification for criteria for evaluating activation and destruction processes of complex ores. *Minerals*, 2023, Vol. 13, No. 5, Art. no. 684. DOI: 10.3390/min13050684

30. **Fedorova E., Pupysheva E., Morgunov V.** Modeling of particle size distribution in the presence of flocculant. *Symmetry*, 2024, Vol. 16, No. 1, Art. no. 114. DOI: 10.3390/sym16010114

31. **Serbin S.D., Smirnova O.A.** Creation of digital twins at mining enterprises. *Tsifrovaia transformatsiia ekonomicheskikh sistem: problemy i perspektivy* (*EKOPROM-2022*) [*Digital transformation of economic systems: problems and prospects* (*ECOPROM-2022*)], 2022, Pp. 541–544. DOI: 10.18720/IEP/2021.4/165

32. **Rozs R., Ando M.** Collaborative systems, operation and task of the manufacturing execution systems in the 21st century industry. *Periodica Polytechnica Mechanical Engineering*, 2020, Vol. 64, No. 1, Pp. 51–66. DOI: 10.3311/PPme.14413

33. **Tishchenko I.V., Vanag Yu.V.** Automation and robotization of solid mineral mining. *Interexpo GEO-Siberia*, 2022, Vol. 2, No. 3, Pp. 325–333. DOI: 10.33764/2618-981X-2022-2-3-325-333

34. **Anufriev A.S., Lebedik E A, Bazhin V.Yu.** New approaches to improving efficiency of automated control over ore pretreatment process stages. *Mining Informational and Analytical Bulletin* (*MIAB*), 2024, Vol 2, Pp. 76–92. DOI: 10.25018/0236_1493_2024_2_0_76

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Bazhin Vladimir Yu.
Бажин Владимир Юрьевич
E-mail: bazhin-alfoil@mail.ru

Anufriev Aleksandr S.
Ануфриев Александр Сергеевич
E-mail: anufriev_rf@yahoo.com

Rusinov Leon A.
Русинов Леон Абрамович
E-mail: lrusinov@yandex.ru

Submitted: 16.09.2024; Approved: 25.03.2025; Accepted: 27.03.2025.
Поступила: 16.09.2024; Одобрена: 25.03.2025; Принята: 27.03.2025.

Research article
DOI: https://doi.org/10.18721/JCSTCS.18203
UDC 004.522,004.934

DATASET CREATION FOR COMPREHENSIVE PERFORMANCE EVALUATION OF AUTOMATIC SPEECH RECOGNITION SYSTEMS

A.Yu. Andrusenko ✉ , *P.D. Drobintsev*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ andrusenkoau@gmail.com

Abstract. The performance evaluation of Automatic Speech Recognition (ASR) systems heavily depends on the availability of diverse and representative test datasets encompassing a wide range of complexities in various domains. This work introduces a novel methodology for collecting and preparing datasets for comprehensive ASR system evaluation. The proposed dataset incorporates a modern vocabulary enriched with numerous unique terms and proper nouns, facilitating an in-depth evaluation of overall ASR performance and the effectiveness of context-biasing techniques in computer science. Additionally, the dataset retains critical text features such as Punctuation and Capitalization (P&C), enabling a rigorous evaluation of P&C prediction algorithms. We present a detailed account of the dataset creation process, along with its statistical and qualitative analysis. Furthermore, we benchmark state-of-the-art ASR models, context-biasing approaches, and P&C prediction techniques using the proposed dataset, providing valuable insights into their relative performance.

Keywords: automatic speech recognition, test dataset, large language models, punctuation and capitalization, context-biasing

Citation: Andrusenko A.Yu., Drobintsev P.D. Dataset creation for comprehensive performance evaluation of automatic speech recognition systems. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 33–44. DOI: 10.18721/JCSTCS.18203

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18203
УДК 004.522,004.934

СОЗДАНИЕ НАБОРА ДАННЫХ ДЛЯ КОМПЛЕКСНОЙ ОЦЕНКИ ПРОИЗВОДИТЕЛЬНОСТИ СИСТЕМ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РЕЧИ

А.Ю. Андрусенко ✉ , *П.Д. Дробинцев*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ andrusenkoau@gmail.com

Аннотация. Оценка производительности систем автоматического распознавания речи (Automatic Speech Recognition, ASR) в значительной степени зависит от наличия разнообразных и репрезентативных тестовых наборов данных, охватывающих широкий спектр сложностей в различных доменах. В данном исследовании представлена новая методология сбора и подготовки наборов данных, предназначенных для всесторонней оценки ASR систем. Предложенный набор данных включает современный словарный запас, обогащенный многочисленными уникальными терминами и именами собственными, что позволяет проводить углубленную оценку общей производительности ASR и эффективности методов смещения контекста (context-biasing) в области компьютерных технологий. Кроме того, в наборе данных сохраняются важные текстовые характеристики, такие как пунктуация и капитализация (Punctuation & Capitalization, P&C), что делает возможной строгую оценку алгоритмов предсказания P&C. Мы подробно описываем процесс создания набора данных, включая его анализ. Более того, мы проводим тестирование передовых ASR моделей, методов смещения контекста и алгоритмов предсказания P&C на основе предложенного набора данных, предоставляя ценные сведения об их относительной производительности.

Ключевые слова: автоматическое распознавание речи, тестовый набор данных, большие языковые модели, пунктуация и капитализация, смещение контекста

Для цитирования: Andrusenko A.Yu., Drobintsev P.D. Dataset creation for comprehensive performance evaluation of automatic speech recognition systems // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 33–44. DOI: 10.18721/JCSTCS.18203

Introduction

The rapid advancements in deep learning techniques have driven the development of numerous end-to-end Automatic Speech Recognition (ASR) systems [1]. A comprehensive evaluation of these models necessitates using test datasets that span a wide range of linguistic and acoustic conditions across diverse domains [2]. While the Word Error Rate (WER) remains the primary metric for assessing overall ASR performance, specific tasks, such as evaluating context-biasing capabilities, are attracting increasing attention. These tasks are designed to measure how effectively an ASR system recognizes domain-specific keywords and terminology [3]. Achieving robust evaluation for context-biasing requires test datasets enriched with novel, domain-specific words, and phrases that may challenge recognition accuracy due to their unfamiliarity.

The most widely used dataset for ASR tasks is LibriSpeech [4], a collection of English audiobook recordings. However, its utility for evaluating high-performance ASR models, such as Whisper [5], is limited due to the dataset's relatively simple data domain. Furthermore, LibriSpeech lacks a substantial number of novel or rare terms, making it unsuitable for evaluating context-biasing capabilities, especially for ASR models trained on extensive datasets.

Other datasets, such as Switchboard [6] and CallHome [7], introduce greater complexity by focusing on conversational telephone speech. Ted-Lium [8], GigaSpeech [9], and People's Speech [10] target ASR evaluation in scenarios resembling YouTube videos and online presentations. Mozilla Common Voice [11] supports other scenarios, featuring dictated, pre-prepared phrases recorded on various personal devices. For more challenging use cases, datasets like AMI [12] and CHiME-5 [13] simulate environments with significant noise, reverberation, and overlapping speakers, presenting additional difficulties for ASR systems. While these datasets allow broader evaluations of model performance, they remain incomplete in their coverage of diverse data domains. Critically, they also lack sufficient quantities of curated keywords and structured lists required for rigorous context-biasing evaluation.

To address the issue of data diversity, the Earnings21/22 [14, 15] public datasets were introduced, featuring earnings calls from nine financial sectors. Alongside the audio data, these datasets include a list of named entities (keywords) designed to facilitate the evaluation of context-biasing techniques. Despite these contributions, the keyword set has notable limitations: it contains many trivial and high-frequency words that most ASR systems already handle effectively, as well as short words (fewer than three characters) that contribute to elevated rates of false acceptance in context-biasing tasks.

Additionally, the dataset lacks segmentation, comprising lengthy audio recordings ranging from five to seventeen minutes. Processing such extended audio sequences can impose significant computational demands on ASR systems, particularly those employing self-attention mechanisms, which often experience out-of-memory issues on GPU hardware during inference.

The ConEC [16] initiative sought to enhance the Earnings21/22 benchmark by segmenting long audio recordings, refining the keyword list, and introducing a publicly available context-biasing solution based on the shallow-fusion decoding approach. However, despite these improvements, the ConEC benchmark remains constrained to a narrow domain focused exclusively on earnings presentations, limiting its applicability for broader ASR evaluation.

This work introduces a novel approach to creating an ASR evaluation dataset, collected from publicly available YouTube channels under a Creative Commons license. The dataset focuses on the modern technology domain, with a particular emphasis on computer science. It features manually annotated transcriptions that preserve Punctuation and Capitalization (P&C), enabling robust evaluation of P&C prediction tasks. Additionally, the dataset includes a diverse set of domain-specific terms, such as product names, making it highly suitable for evaluating context-biasing methods. To further support these evaluations, we also propose a method for generating keyword lists tailored to the context-biasing task.

The dataset preparation process is implemented using open-source tools within the NeMo framework[1]. The preparation pipeline incorporates several key stages: text cleaning and normalization, automated punctuation insertion using a Large Language Model (LLM), segmentation of data through the removal of non-speech segments, and additional filtering based on ASR accuracy thresholds. These steps ensure a high-quality and domain-relevant dataset for ASR evaluation.

We conducted experiments on the proposed dataset using state-of-the-art ASR models from Hugging Face. Our evaluations included assessments of overall ASR performance, the effectiveness of context-biasing techniques using the proposed keyword list, and P&C prediction accuracy. The results provide valuable insights into the capabilities and limitations of the evaluated models within this domain-specific dataset.

Data preparation pipeline

To enhance ASR evaluation benchmarks under modern conditions, we focused on data scenarios relevant to the field of computer science. A prime example of such data is keynote presentations on various technology topics from major tech companies, such as Google, Microsoft, Amazon, and others. Using

[1] GitHub – NVIDIA/NeMo: A scalable generative AI framework built for researchers and developers working on Large Language Models, Multimodal, and Speech AI (Automatic Speech Recognition and Text-to-Speech), Available: https://github.com/NVIDIA/NeMo (Accessed 21.05.2025)

the YouTube-dl library[2], we collected 15 hours of full-length recordings from recent years, capturing content directly from these events.

The collected recordings include manually created transcriptions with preserved P&C, making them valuable for tasks, such as P&C prediction. However, the raw data required extensive preprocessing to ensure its suitability for ASR evaluation benchmarks.

Text preprocessing

Even manually created transcriptions can contain numerous typos and non-standard characters, negatively impacting ASR evaluation. To address this, we applied pattern-based substitutions using regular expressions to correct common errors and remove invalid characters.

Text normalization was performed to convert numerical values and auxiliary symbols into their text representations. This process was implemented using the NeMo Text Processing toolkit[3], which supports both forward and inverse text normalization. The normalization process ensured that only characters from the English alphabet were retained in the processed dataset. Additionally, NeMo Text Processing supports audio-based text normalization, which leverages baseline ASR model outputs to enhance numeral normalization. While this method can improve accuracy, it has the potential to introduce errors in challenging acoustic conditions due to ASR recognition inaccuracies.

For punctuation, we standardized the dataset to include only three primary punctuation marks: periods, commas, and question marks. This simplification ensures consistency while maintaining sufficient information for P&C tasks.

Punctuation reconstruction with LLM

Certain portions of the collected data lacked P&C. To address this, we employed the Llama-3-8B[4] LLM, utilizing a carefully designed prompt. The prompt included standardized instructions: "Your task is to punctuate the input text. You can only use a period, comma, or question mark as punctuation. Add capitalization to the beginning of new sentences."

To process entire text files, we adopted a chunk-based approach, dividing the text into segments of 250 words per iteration. However, this method introduced a potential issue: chunks could end mid-sentence, leading the LLM to erroneously assign an end-of-sentence punctuation mark (e.g., a period or question mark) to the last word in the chunk. To mitigate this, we extracted only the first n-1 complete sentences from each processed chunk, avoiding disruptions caused by mid-sentence breaks. The subsequent chunk then began at the last valid sentence boundary of the previous segment.

While the LLM effectively added punctuation and restored missing capitalization, it slightly altered the original text. In our evaluation, the WER between the input transcription and the normalized LLM output was approximately 2%. This discrepancy was primarily due to the LLM's removal of repetitive words typical in spoken language and its attempts to correct typos introduced during manual transcription. These issues suggest potential improvements with prompt refinement.

Table 1 illustrates an example of text correction during punctuation restoration using the Llama-3-8B model.

Segmentation

The original dataset comprises full-length recordings ranging from 1 to 2 hours. Such lengthy inputs pose challenges for ASR systems utilizing global attention mechanisms, as their quadratic complexity with respect to input sequence length can lead to significant computational overhead. Additionally, the presence of prolonged musical segments in the recordings may degrade speech recognition accuracy.

Although the source data includes original timestamps associated with the corresponding text (subtitles), direct segmentation based on these timestamps often results in errors at segment boundaries,

[2] GitHub – ytdl-org/youtube-dl: Command-line program to download videos from YouTube.com and other video sites, Available: https://github.com/ytdl-org/youtube-dl (Accessed 21.05.2025)

[3] GitHub – NVIDIA/NeMo-text-processing: NeMo text processing for ASR and TTS, Available: https://github.com/NVIDIA/NeMo-text-processing (Accessed 21.05.2025)

[4] meta-llama/Meta-Llama-3-8B-Instruct · Hugging Face, Available: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct (Accessed 21.05.2025)

Table 1

Example of original transcription and Llama-3-8B punctuation reconstruction. Reference typos, LLM corrections, and removed words are highlighted in red and green colors

before	we thought it was a perfect perfect name **black wealth**
after	We thought it was a perfect perfect name, **Blackwell**.
before	..this software and it **compress** it dimensionally **reduce** it..
after	..this software and it **compresses** it, dimensionally **reduces** it..
before	..the new **micr service** now the thing that's that's emerging here..
after	..the new **microservice**. Now, the thing that's emerging here..

thereby increasing the WER of the segmented dataset. Consequently, we utilized these timestamps solely to remove extended non-speech segments at the beginning and end of the audio files.

For more accurate data segmentation, we employed CTC-segmentation [17] using the NeMo toolkit. This method aligns ground truth transcriptions with corresponding audio files effectively. We used Citrinet [18], a convolutional neural network ASR model, for segmentation. Citrinet is particularly well-suited for handling long audio files without encountering GPU memory limitations. Furthermore, its CNN-based architecture ensures more precise alignments, avoiding the late or early prediction errors common in attention-based models.

To streamline the segmentation process, all text data was initially divided into individual sentences based on punctuation. This allowed us to determine sentence boundaries and obtain alignment confidence scores for each segment. However, short sentences often exhibited boundary errors during alignment (Fig. 1, *a*). To mitigate these issues, we merged consecutive sentences if the silence between them was less than one second and their alignment confidence score exceeded –5.0. This approach produced final audio segments with durations ranging from 2 to 40 seconds (Fig. 1, *b*).

This refined segmentation strategy reduced the WER from 12.68% to 11.78% on the processed dataset. However, it also shifted the duration distribution toward the upper limit of 40 seconds, reflecting a bias toward longer segment lengths.

To enhance the diversity of segment lengths, we implemented a probabilistic sentence merging approach. Instead of enforcing mandatory sentence merging up to a predefined length threshold (based on the previously described conditions), we applied a probabilistic mechanism. Specifically, each subsequent sentence was merged with the current segment with a probability of 0.8, provided that the conditions for silence duration and confidence score were satisfied.

The outcomes of this probabilistic sentence merging approach are illustrated in fig. 1c. This method successfully increased the diversity of segment lengths while maintaining comparable ASR performance, as no significant degradation in WER was observed.

Data filtering

Manual analysis of the recognition results revealed that examples with high WER were predominantly caused by segmentation errors or inaccuracies in the reference transcriptions. To address this, we applied a filtering process based on ASR performance metrics.

As a baseline ASR model, we utilized the Fast Conformer-Transducer Large (114M parameters)[5], trained on 20,000 hours of English speech data. We filtered out examples where the WER exceeded 80%, or the Character Error Rate (CER) exceeded 30%. This process resulted in a cleaner, segmented evaluation dataset with a total duration of 12.4 hours.

[5] nvidia/stt_en_fastconformer_ctc_large · Hugging Face, Available: https://huggingface.co/nvidia/stt_en_fastconformer_ctc_large (Accessed 15.06.2024)

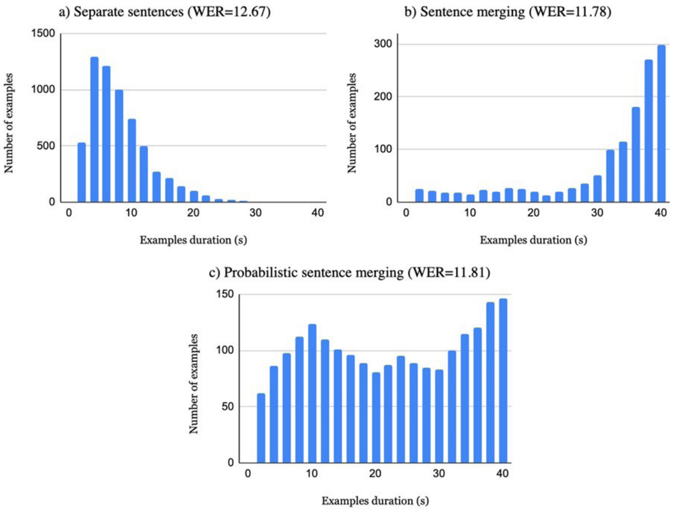


Fig. 1. Duration distribution of segmented data according to the different merge methods: separate sentences, sentence merging, and probabilistic sentence merging. WER of the data sets obtained by considered segmentation methods is 12.67%, 11.78%, and 11.81%, respectively

To prepare for the evaluation, we divided the obtained dataset into two subsets: a 4-hour development (dev) set and an 8.4-hour test set, ensuring non-overlapping talks between the two subsets. All subsequent evaluation results are reported exclusively for the test set.

Named entities (keywords)

The proposed dataset includes a substantial number of named entities suitable for context-biasing tasks. To analyze Named Entity Recognition (NER) statistics, we used SpaCy[6], following a similar methodology to previous works. SpaCy assigns entity tags to words based on predefined classes, such as ORG (organization), PERSON (person), DATE, and CARDINAL (numbers). The proposed dataset contains a significant number of examples for these tags (e.g., ORG=3,534, CARDINAL=1,457, PERSON=846, DATE=987, etc.).

However, SpaCy's tagging process introduces challenges, including overlapping classifications (e.g., the word "AI" being tagged as both ORG and PERSON) and classification errors. Additionally, most words in this entity list achieve high recognition accuracy, when evaluated using the baseline ASR model, making them less relevant for assessing context-biasing performance. For our analysis, we focused on identifying words with low recognition accuracy.

To construct a more appropriate context-biasing keyword list, we applied the following methodology:

1. ASR Evaluation: The dataset was transcribed using the baseline ASR model, and recognition accuracy was calculated for individual words (monograms) and phrases (bigrams).
2. Entity Filtering: Only words present in the named entities identified by SpaCy were retained.
3. Error Word Identification: We observed that most misrecognized phrases contained a single error word already represented in the monogram statistics. Consequently, we prioritized individual words over bigrams, selecting only a limited number of bigrams.
4. Short Word Exclusion: Words shorter than three characters were excluded, as these often contribute to high false acceptance rates during context-biasing recognition.

This process resulted in a refined list of 200 keywords with low recognition accuracy, suitable for evaluating context-biasing techniques. Additionally, we incorporated 800 distractor words – terms likely absent from the dataset – sourced from the Earnings benchmark. This combination allows for a more rigorous evaluation of context-biasing performance.

[6] spaCy · Industrial-strength Natural Language Processing in Python, Available: https://spacy.io (Accessed 21.05.2025)

Experimental setup

Speech recognition

To assess speech recognition accuracy, measured by WER, on the obtained dataset, we evaluated a selection of top-performing public models listed on the Hugging Face ASR Leaderboard[7]. This leaderboard ranks ASR models based on their average WER across multiple public test sets and includes metrics for inference speed. At the time of evaluation, the leading models included those from the NeMo toolkit (e.g., Conformer, Fast-Conformer, Parakeet, and Canary) and OpenAI (Whisper-large-v1/v2/v3).

To ensure fair comparison across models, we applied consistent normalization to all recognition outputs, following the data preparation procedure. This included expanding numerical symbols into their textual representations, removing punctuation, and converting all text to lowercase.

Punctuation and capitalization

To evaluate the capabilities of ASR models in P&C prediction, we selected public models that inherently support P&C functionality. From the NeMo toolkit, we chose the three highest-performing models with P&C capabilities at the time: Fast-Conformer Hybrid with P&C (operating in Transducer decoding mode), Parakeet-tdt_ctc-1.1b, and Canary-1b. Similarly, we selected the top three models from OpenAI's Whisper series (Whisper-large-v1/v2/v3). All these models are available on the Hugging Face platform.

To measure P&C performance, we used two key metrics:

- WER C – Word Error Rate calculated with capitalization preserved in the text.
- PER – Punctuation Error Rate, focusing exclusively on punctuation errors:

$$\mathrm{PER} = \frac{\mathrm{I} + \mathrm{D} + \mathrm{S}}{\mathrm{I} + \mathrm{D} + \mathrm{S} + \mathrm{C}}, \tag{1}$$

where I, D, S, and C are the number of insertions, deletions, substitutions, and correct punctuation predictions during a backtrace matrix calculation. More details about WER C and PER metrics can be found in [19].

Context-biasing

To assess context-biasing performance, we explored the available techniques from the NeMo toolkit using a Hybrid Transducer-CTC model[8]. Notably, this model was not trained on data from the computer science domain, ensuring a fair evaluation of context-biasing methods.

One method to enhance keyword recognition accuracy is word boosting, supported via pyctcdecode[9]. This technique employs a shallow fusion approach during the CTC beam-search decoding process. We applied the default parameters, setting the keyword boosting weight (hotword_weight = 10) and beam size (beam_size = 5).

Another approach is the fast context-biasing method using the CTC-based Word Spotter (CTC-WS) [19]. This method decouples keyword recognition from the broader recognition process by leveraging fast decoding of CTC logits based on a graph constructed exclusively from context-specific keywords. The Word Spotter identifies the desired keywords along with their time intervals and confidence scores, which are subsequently merged with the results from greedy decoding. When used with a Hybrid CTC-Transducer model, this technique enables keyword boosting within Transducer predictions. For this evaluation, we applied the default boosting parameters.

[7] Open ASR Leaderboard – a Hugging Face Space by hf-audio, Available: https://huggingface.co/spaces/hf-audio/open_asr_leaderboard (Accessed 21.05.2025)

[8] STT En FastConformer Hybrid Transducer-CTC Large Streaming Multi | NVIDIA NGC, Available: https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_fastconformer_hybrid_large_streaming_multi (Accessed 21.05.2025)

[9] GitHub – kensho-technologies/pyctcdecode: A fast and lightweight python-based CTC beam search decoder for speech recognition, Available: https://github.com/kensho-technologies/pyctcdecode (Accessed 21.05.2025)

As metrics for context-biasing evaluation, we used the standard WER for the entire text and F-score for words from the context-biasing list:

$$F_{score} = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

In addition, we measured the speed performance of the considered methods excluding the encoder work time. All runtime measurements are averaged over 10 runs.

Experimental results

Speech recognition

Table 1 presents the WER performance of the previously discussed ASR models on both the Hugging Face Leaderboard (HF LB) test sets and the proposed dataset (first two columns). The results indicate that the top-performing models on the HF LB also exhibit strong performance on the proposed test set, achieving WER values of approximately 6–7%. This high accuracy may be attributed to the extensive training data utilized by these models, which likely includes diverse sources such as YouTube content.

The best WER on the proposed dataset set was achieved by the Canary-1b model, which was also the top-ranked model on the HF LB at the time of evaluation.

Table 2

Performance results (%) of public ASR models from Hugging Face

ASR model	Size, B	HF LB, WER	Proposed dataset		
			WER	WER C	PER
C_transducer_L	0.12	10.2	15.4	–	–
FC_transducer_L	0.12	9.8	12.1	–	–
FC_hybrid_L_pc	0.12	–	9.4	13.3	37.5
Parakeet-tdt_ctc	1.10	8.1	6.7	10.1	25.4
Canary-1b	1.00	7.7	6.4	9.7	26.3
Whisper-large-v1	1.55	10.4	6.5	10.3	27.1
Whisper-large-v2	1.55	9.0	6.8	9.6	27.4
Whisper-large-v3	1.55	8.6	6.6	9.5	27.5

Note: HF LB WER is the average WER for other test sets in the HF LB; WER C is the WER with capitalization left in the text; PER is Punctuation Error Rate

However, smaller models (about 120M parameters) based on Conformer and Fast-Conformer architectures with only public training datasets (20k+ hours) showed WER above 10% for the proposed test set. This fact confirms the absence of the proposed computer science domain in public datasets.

Punctuation and capitalization

The results for P&C prediction are presented in the rightmost columns of Table 2, which include metrics for WER C and PER. While these metrics correlate with the standard WER, discrepancies can arise in specific cases. For instance, when comparing the Parakeet and Canary models, a model with a better WER may perform worse in terms of PER. This highlights the importance of evaluating punctuation and capitalization effectiveness independently, enabled by datasets that preserve P&C information, in addition to standard transcription accuracy.

An ablation study on punctuation prediction is illustrated in Fig. 2. This analysis examines the prediction error rates for individual punctuation marks: periods, commas, and question marks across the

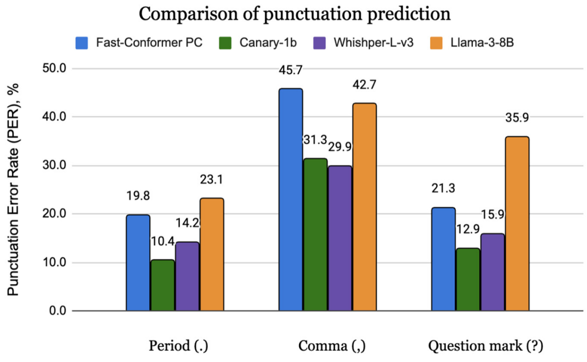


Fig. 2. Comparison of punctuation prediction (PER) for three considered punctuation classes over the proposed test set except samples with Llama-3-8B punctuation. The overall PER is presented in the legend

three evaluated ASR models. Additionally, the study includes a comparison with punctuation predictions generated by the Llama-3 LLM, which was employed during the dataset preparation process. To ensure fairness in comparison, test set examples containing punctuation generated by the LLM were excluded from the evaluation.

The results demonstrate that commas have the lowest prediction accuracy compared to periods and question marks, which are recognized with relatively high accuracy. This finding highlights the inherent difficulty of predicting commas in ASR systems. As a potential improvement, it may be preferable to omit comma prediction and focus solely on sentence-end punctuation, such as periods and question marks.

Punctuation accuracy achieved by the LLM model was lower than that of the top-performing ASR models. This discrepancy underscores the advantage of leveraging audio cues, which significantly enhance punctuation accuracy, particularly for question marks. Nevertheless, the LLM performed reasonably well when relying solely on textual input, making it a viable option for simplifying punctuation tasks in long audio files. Using ASR models for punctuation in such scenarios can be computationally intensive, as it requires chunk-wise decoding, alignment of ASR outputs with the original text, and the subsequent merging of results. The LLM offers a simpler alternative for such use cases.

An additional analysis investigated how the number of sentences within test examples affects the accuracy of end-of-sentence punctuation predictions. For single-sentence test examples, this task is relatively straightforward, as the model can predict sentence-end punctuation with high confidence at the end of the input. However, for test examples containing multiple sentences, the task becomes increasingly complex.

To quantify this effect, we measured the PER for sentence-end labels (periods and question marks) across all test examples. We then grouped the results based on the number of sentences in the reference transcriptions and averaged the PER for each group. The findings, presented in Fig. 3, confirm the hypothesis: as the number of sentences in test examples increases, the accuracy of sentence-end punctuation predictions decreases. This observation supports the utility of grouping multiple sentences into single test examples to increase the overall complexity of punctuation evaluation tasks.

Context-biasing

The results of the context-biasing evaluation for the proposed dataset are summarized in Table 3. Both context-biasing methods demonstrated improvements in recognition accuracy. However, pyctcdecode exhibited relatively poor performance compared to the CTC-WS method.

One significant limitation of pyctcdecode is its sensitivity to the size of the context-biasing list. Due to a marked degradation in processing speed with larger lists, we were constrained to use a list of 200 words instead of the initially intended 1000 words. This limitation aligns with observations from prior

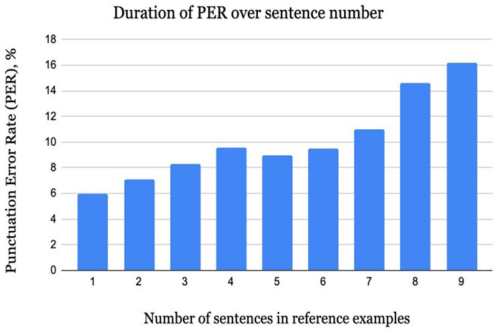


Fig. 3. PER distribution for the period and the question mark over a number of sentences in reference examples in the proposed test set. Canary-1b was used as an ASR model

work, which similarly reported performance issues when scaling the context-biasing list size in pyctcdecode.

The CTC-WS method, when applied with the proposed context-biasing list, improved recognition accuracy for both CTC and Transducer decodings by over 4%, with minimal additional decoding time overhead. These results highlight the abundance of keywords in the proposed dataset, making it highly suitable for evaluating context-biasing tasks, particularly in scenarios involving novel domains.

Table 3

CTC and Transducer decoding results for the proposed test set

Method	CB	Time, s	F-score (P/R)	WER, %
CTC				
greedy	no	4	0.36 (0.96/0.21)	16.44
pyctcdecode	no	21	0.37 (0.97/0.23)	16.57
	yes	1498	0.66 (0.87/0.54)	15.41
CTC-WS	yes	31	0.82 (0.82/0.82)	12.07
Transducer				
greedy	no	15	0.41 (0.97/0.26)	15.89
CTC-WS	yes	44	0.82 (0.82/0.82)	11.69

Note: CB is the presence of context-biasing; P is Precision; R is Recall.

Overall assessment of the proposed methodology

Based on the conducted analysis of the ASR systems evaluation, we can draw a conclusion about the effectiveness of the proposed methodology. For example, the use of LLM allows us to arrange and evaluate the accuracy of P&C recognition, which is also important when segmenting long audios by sentences. Probabilistic sentence merging allows us to simultaneously reduce the number of segmentation errors at the edges of segments and make examples with several sentences that improve P&C evaluation. The choice of keywords allows us to test various context-biasing methods, which are extremely important at this time. The proposed methodology for collecting and processing data allows us to obtain a versatile high-quality test set for broad performance evaluation of modern ASR systems in three main areas.

Conclusion

This study introduced a novel methodology for preparing evaluation datasets tailored to the computer science domain. The resulting dataset features a rich set of domain-specific terms and retains punctuation and capitalization (P&C), enabling its use for comprehensive ASR model evaluation. It supports a broad range of tasks, including standard speech recognition (WER), context-biasing, and P&C prediction scenarios.

We provided a detailed description of the data preparation pipeline, utilizing open-source frameworks to ensure reproducibility and accessibility. Furthermore, we evaluated state-of-the-art public ASR models across the three primary use cases and conducted ablation studies on punctuation prediction using the proposed dataset. This work offers a robust resource and valuable insights for advancing ASR evaluation benchmarks.

REFERENCES

1. **Prabhavalkar R., Hori T., Sainath T.N., Schlüter R., Watanabe S.** End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024, Vol. 32, Pp. 325–351. DOI: 10.1109/TASLP.2023.3328283

2. **Nguyen T.S., Stüker S., Waibel A.** Toward cross-domain speech recognition with end-to-end models. *arXiv:2003.04194*, 2020. DOI: 10.48550/arXiv.2003.04194

3. **Pundak G., Sainath T.N., Prabhavalkar R., Kannan A., Zhao D.** Deep context: End-to-end contextual speech recognition. *arXiv:1808.02480*, 2018. DOI: 10.48550/arXiv.1808.02480

4. **Panayotov V., Chen G., Povey D., Khudanpur S.** Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2015, Pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964

5. **Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I.** Robust speech recognition via large-scale weak supervision. *arXiv:2212.04356*, 2022. DOI: 10.48550/arXiv.2212.04356

6. **Godfrey J.J., Holliman E., McDaniel J.** SWITCHBOARD: telephone speech corpus for research and development. *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, Vol. 1, Pp. 517–520. DOI: 10.1109/ICASSP.1992.225858

7. **Cieri C., Miller D., Walker K.** The fisher corpus: A resource for the next generations of speech-to-text. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (*LREC'04*), 2004.

8. **Rousseau A., Deléglise P., Estève Y.** TED-LIUM: an automatic speech recognition dedicated corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (*LREC'12*), 2012, Pp. 125–129.

9. **Chen G., Chai S., Wang G., Du J., Zhang W.-Q., Weng C., Su D., Povey D., Trmal J., Zhang J., Jin M., Khudanpur S., Watanabe S., Zhao S., Zou W., Li X., Yao X., Wang Y., You Z., Yan Z.** GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. *INTERSPEECH 2021*, 2021, Pp. 3670–3674. DOI: 10.21437/Interspeech.2021-1965

10. **Galvez D., Diamos G., Ciro J., Cerón J.F., Achorn K., Gopi A., Kanter D., Lam M., Mazumder M., Reddi V.J.** The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *35th Conference on Neural Information Processing Systems* (*NeurIPS 2021*), 2021, Pp. 1–12.

11. **Ardila R., Branson M., Davis K., Henretty M., Kohler M., Meyer J., Morais R., Saunders L., Tyers F.M., Weber G.** Common voice: A massively-multilingual speech corpus. *arXiv:1912.06670*, 2019. DOI: 10.48550/arXiv.1912.06670

12. **Carletta J., Ashby S., Bourban S. et al.** The AMI meeting corpus: A pre-announcement. *Machine Learning for Multimodal Interaction* (*MLMI 2005*), 2005, Vol. 3869, Pp. 28–39. DOI: 10.1007/11677482_3

13. **Barker J., Watanabe S., Vincent E., Trmal J.** The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, 2018.

14. **Del Rio M., Delworth N., Westerman R., Huang M., Bhandari N., Palakapilly J., McNamara Q., Dong J., Żelasko P., Jetté M.** Earnings-21: A practical benchmark for ASR in the wild. *INTERSPEECH 2021*, Pp. 3465–3469. DOI: 10.21437/Interspeech.2021-1915

15. **Del Rio M., Ha P., McNamara Q., Miller C., Chandra S.** Earnings-22: A practical benchmark for accents in the wild. *arXiv:2203.15591*, 2022. DOI: 10.48550/arXiv.2203.15591

16. **Huang R., Yarmohammadi M., Trmal J., Liu J., Raj D., Garcia L.P., Ivanov A.V., Ehlen P., Yu M., Povey D., Khudanpur S.** ConEC: Earnings call dataset with real-world contexts for benchmarking contextual speech recognition. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (*LREC-COLING 2024*), 2024, Pp. 3700–3706.

17. **Kürzinger L., Winkelbauer D., Li L., Watzel T., Rigoll G.** CTC-segmentation of large corpora for German end-to-end speech recognition. *Speech and Computer*, 2020, Pp. 267–278. DOI: 10.1007/978-3-030-60276-5_27

18. **Majumdar S., Balam J., Hrinchuk O., Lavrukhin V., Noroozi V., Ginsburg B.** Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. *arXiv:2104.01721*, 2021. DOI: 10.48550/arXiv.2104.01721

19. **Meister A., Novikov M., Karpov N., Bakhturina E., Lavrukhin V., Ginsburg B.** LibriSpeech-PC: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end ASR models. *arXiv:2310.02943*, 2023. DOI: 10.48550/arXiv.2310.02943

20. **Andrusenko A., Laptev A., Bataev V., Lavrukhin V., Ginsburg B.** Fast context-biasing for CTC and transducer ASR models with CTC-based word spotter. *INTERSPEECH 2024*, 2024, Pp. 757–761. DOI: 10.21437/Interspeech.2024-1002

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Andrusenko Andrei Yu.
Андрусенко Андрей Юрьевич
E-mail: andrusenkoau@gmail.com

Drobintsev Pavel D.
Дробинцев Павел Дмитриевич
E-mail: drob@ics2.ecd.spbstu.ru
ORCID: https://orcid.org/0000-0003-1116-7765

Submitted: 18.12.2024; Approved: 17.04.2025; Accepted: 30.04.2025.
Поступила: 18.12.2024; Одобрена: 17.04.2025; Принята: 30.04.2025.

Computing, Telecommunication and Control, 2025, Vol. 18, No. 2, Pp. 45–55.
Информатика, телекоммуникации и управление. 2025. Том 18, № 2. С. 45–55.

Research article
DOI: https://doi.org/10.18721/JCSTCS.18204
UDC 004.42

A PAGE-BASED APPROACH FOR STORING VECTOR EMBEDDINGS

N.A. Tomilov ✉ , *V.P. Turov*

ITMO University, Saint Petersburg, Russian Federation

✉ firemoon@icloud.com

Abstract. This study proposes a page-based approach to organize the storage for vector embeddings combined with the use of general-purpose lossless compression algorithms. The proposed approach organizes vector embeddings into pages of a configurable number of entries that contain vector embeddings and all necessary metainformation, and then the page files are compressed using general-purpose compression algorithms. This approach allows configuring page size and specific compression algorithm, to balance retrieval speed and storage efficiency. Experiments on three datasets, including PyEmb-50GB with more than 28 million dense vector embeddings, showed that the proposed solution reduces the occupied disk space by 14–40% compared to existing storage formats, such as ORC and Parquet, and up to two times compared to SQLite and H2. In addition, the suggested approach demonstrates a comparable to SQLite and H2 vector retrieval time, which is also a hundred times faster than ORC and Parquet. The results indicate that increasing the page size logarithmically reduces the storage size, while linearly increasing retrieval time. The proposed storage format supports thread-safe vector access, reducing both the necessary disk space and retrieval time, making it a robust solution for large-scale vector data management. It can also be used in approximate nearest neighbor search, provided the correct way of sharding vector embeddings between pages.

Keywords: vector embeddings, compression of vector embeddings, ORC, Parquet

Citation: Tomilov N.A., Turov V.P. A page-based approach for storing vector embeddings. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 45–55. DOI: 10.18721/JCSTCS.18204

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18204
УДК 004.42

ПОДХОД ДЛЯ СТРАНИЧНОЙ ОРГАНИЗАЦИИ ХРАНЕНИЯ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ

Н.А. Томилов ✉ , *В.П. Туров*

Университет ИТМО, Санкт-Петербург, Российская Федерация

✉ firemoon@icloud.com

Аннотация. В данном исследовании предложен страничный подход к организации хранения векторных представлений в сочетании с использованием универсальных алгоритмов сжатия без потерь. Предложенный подход организует векторные представления в страницы из конфигурируемого числа записей, хранящих векторные представления и необходимую метаинформацию, после чего сжимает файлы страниц алгоритмами сжатия общего назначения. Такой подход позволяет задавать настраиваемый размер страницы и выбирать необходимый алгоритм сжатия, обеспечивая баланс между скоростью извлечения данных и эффективностью использования дискового пространства. Эксперименты на трех наборах данных, включая PyEmb-50GB с более чем 28 миллионами плотных векторных представлений, показали, что предложенное решение уменьшает занимаемый объем дискового пространства на 14–40% по сравнению с существующими форматами хранения, такими как ORC и Parquet, и до двух раз по сравнению с SQLite и H2. Помимо этого, предложенное решение демонстрирует сопоставимое с SQLite и H2 и на два порядка меньшее по сравнению с ORC и Parquet время извлечения векторного представления. Результаты демонстрируют, что увеличение размера страницы логарифмически снижает объем хранилища, при этом время извлечения данных увеличивается линейно. Предложенный формат хранения обеспечивает потокобезопасный доступ к векторным представлениям, уменьшая занимаемое дисковое пространство и время доступа. Это делает его надежным решением для управления большими объемами векторных данных. Формат также может быть использован для задач поиска приблизительных ближайших соседей при корректном распределении векторных представлений по страницам.

Ключевые слова: векторные представления, сжатие векторных представлений, ORC, Parquet

Для цитирования: Tomilov N.A., Turov V.P. A page-based approach for storing vector embeddings // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 45–55. DOI: 10.18721/JCSTCS.18204

Introduction

Humanity generates vast amounts of data in various formats and requires rapid access to this information. Machine learning-based search algorithms have gained significant popularity, as they create representations that capture the semantic structure of both textual [1] and multimodal documents in the form of vector embeddings – sequences of floating-point numbers. Thus, the information retrieval process is organized by performing operations on these vectors [2]. Our previous work explored reducing data storage requirements through scalar quantization, resulting in lossy compression, followed by clustering and further quantization [3]. As the term itself suggests, lossy compression, such as quantization, transforms a vector embedding into a more compact representation that is still suitable for machine learning tasks [4]. However, such transformation comes at the cost of precision, which negatively impacts search quality metrics compared to the original dataset [5]. This approach becomes unsuitable when exact vector search is required, rather than an approximate nearest neighbor search. Preserving vector embeddings in their original form demands significant disk space, motivating the use of lossless compression and encoding techniques.

To store large volumes of data, specialized serialization data formats and libraries are often used, with Apache ORC and Apache Parquet being the most prominent examples [6]. These libraries store data as records characterized by predefined fields and field types. The records are divided into smaller chunks and saved to storage, often using various compression algorithms. However, these libraries have significant drawbacks, primarily the lack of support for random access to data. Retrieving a record by a specific index is possible only with a query and additional tools, such as bloom filters or dictionaries, which could result in iterating over multiple records to find the necessary one, slowing down the access time [7].

We propose an approach to storing vector embeddings as collections of files, called pages, each containing a fixed number of vector embeddings. These pages are indexed to store the offsets of vector embeddings belonging to the original documents. Our hypothesis suggests that such paginated storage of vector embeddings will enable more efficient data compression compared to compressing embeddings individually, while achieving a balance between compression efficiency and minimal retrieval time for individual embeddings.

Page-based approach for storing vector embeddings

In this and our previous work, we define a vector embedding to be represented by an array of floating-point numbers, identified by a primary identifier, also called a document index, which maps the vector embedding to the original document from which this vector embedding was acquired. An additional metadata can also be present for vector embeddings, such as a secondary identifier, that maps the embedding to a specific section or sentence in the document, or a cluster identifier of the embeddings based on the document, in case a single original document was turned into multiple embeddings [8].

The essence of the approach lies in grouping and serializing vector embeddings on disk as specified below. First, the original set of vector embeddings is grouped into M groups based on a certain rule being a hyper-parameter of this approach. As an example of such a rule, it could be k-means clustering [9], meaning the groups are resulting clusters, or a rule, according to which the number of vectors in a target group does not exceed a certain threshold. In this research, we use the latter with the threshold value N.

Then, the embeddings belonging to the same original document are grouped together to form an entry. Entries are serialized to byte arrays, so that such serialized entries contain all the necessary metadata, apart from the document index, which is shared across all embeddings within the entry, as well as all vector embedding values represented as 32-bit floating-point numbers. All serialized entries within the group form a page. The page is then written to a page file on disk. A secondary file for a page, called page index, is also written on disk. This page index contains a list of pairs that map the document index to an offset on the disk where the entries with this document index are stored.

Finally, a page file, being a plain binary file, is compressed using any general-purpose compression algorithm. This results in the necessity to decompress the entire page file to access vector embedding, which increases the access time, however, this drawback is offset by lowering storage space necessary to store the data.

The steps described above are shown in Fig. 1.

According to the rule mentioned above, each storage page can contain no more than N vector embeddings, where N is specified during the storage creation. If the number of embeddings associated with a particular primary identifier exceeds N, they are divided into multiple entries distributed across different storage pages. It means that a single primary identifier may appear in multiple entries across various pages. However, within a single page, each primary identifier corresponding to an entry is unique. The hierarchy between the embeddings, entries and pages is represented in Fig. 2.

Accessing a vector embedding for a given document index involves two stages. In the first stage, all page index files are scanned sequentially to find the necessary entry. If there is such a possibility, those page index files can be loaded into memory beforehand to lower the access time. If an index file contains the target primary identifier, the corresponding entry is retrieved from the page: the page data file is

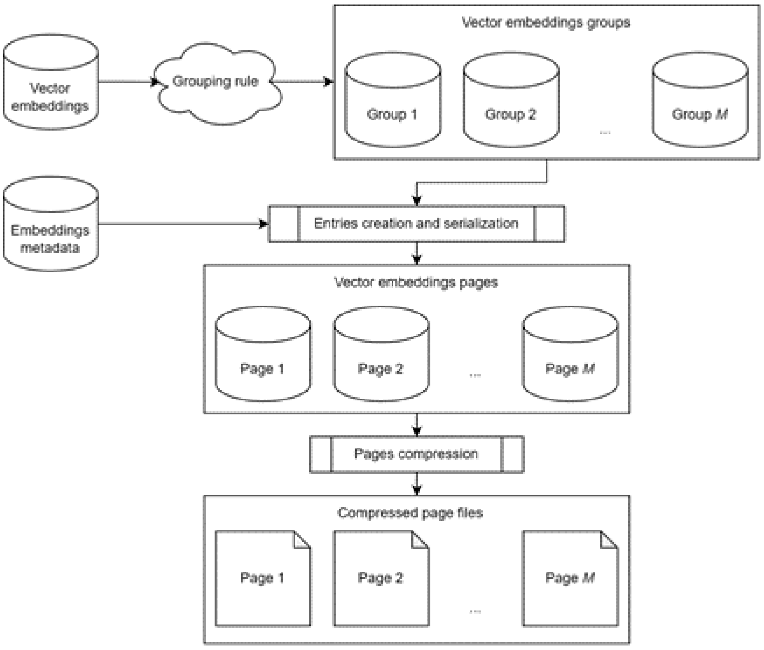


Fig. 1. Process of grouping embeddings to form page files

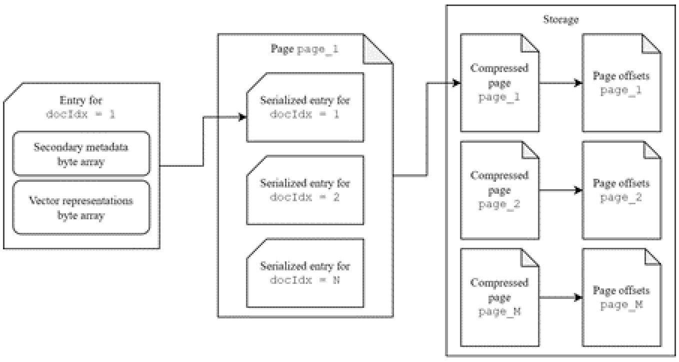


Fig. 2. A structure and relation of entries and pages

decompressed, and the entry is read starting from the specified byte offset. Since the page index file stores tuples containing a primary identifier with byte offset in the data file, ordered by byte offset, the entry size is determined as the difference between the offsets of the next and current entry. Then, the necessary vectors are fetched from the entry. If a more complex query is necessary, like fetching by primary and secondary index, an additional filtration of embeddings within the entry will be necessary.

Modification or deletion of embeddings or their metadata is not supported. However, it is possible to create a copy of the page, excluding the data that should be deleted or modified, while adding new, modified data. This lowers the applicability of this approach to use only for long-term storage.

Since pages are independent from each other, it is possible to operate concurrently over multiple pages with multiple threads, while ensuring that a single thread is performing writing operations over a single page at a time. Concurrent reading operations over a single page by multiple entries is possible, allowing for multithreaded full traversals, similar to a full table scan operation [10] in relational databases. This is crucial for exporting data to other storage systems or performing exact nearest-neighbor vector searches.

Benefits and drawbacks

The main difference between the proposed storage implementation and well-known serialization formats, such as Parquet and ORC, is its ability to improve the random access to the entry by its index. Even though both Parquet and ORC support indexing, it is limited [6] and it did not work reliably in our experiments. Another key benefit is the use of the grouping rule. In this work, we focused on grouping vector embeddings to pages just by the number of vectors per page. However, as stated above, there could potentially be other mechanisms of grouping, for example, using k-means clustering. Such clustering makes the approximate nearest neighbor search possible, reducing the full scan of the entire storage to the full scan of the pages closest to the search query [11], making the proposed solution suitable for systems requiring such search.

The main drawback of this approach is the number of separate files on the filesystem. Each page is stored as a couple of separate files, which means that in instances with a large number of small pages the overhead of storing small files will be significant. This could be mitigated by merging multiple pages into a single file, but it was beyond the scope of this research.

Another drawback is the access slowdown caused by decompressing page files on every access, which could be mitigated by having a decompressed cache of the most or least used pages. It is worth mentioning that the proposed solution describes only a storage layer and cannot serve as a dedicated vector database without most of the features associated with these databases, such as remote access, and thus cannot replace or compare with databases, such as Milvus or Pinecone. Instead, implementing this approach results in replacing the built-in application-level storage, such as the SQLite or H2.

Approach implementation

To test the viability of the proposed approach, it was implemented in Kotlin programming language as a storage library for use in the JVM ecosystem. The data in the pages is serialized into a byte array using the Apache Avro format. As the additional metadata a secondary identifier was chosen. The Avro schema that is used to serialize individual entries is presented in Fig. 3. Such entries are then joined into page files and compressed as explained above.

Run-length encoding [12] is used to serialize page offsets. Several compression algorithms, such as deflate [13], LZMA, LZMA2 [14] and ZStd [15] were chosen, because they are provided by Apache Commons libraries, and were configured to maximize the compression ratio. The maximum page size N and lossless compression algorithm is configurable during storage creation.

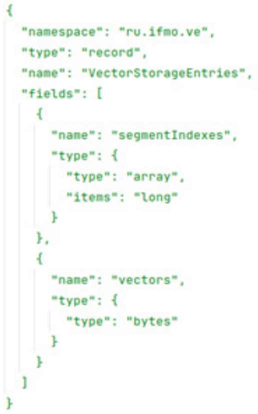
Experiment setup

To evaluate the proposed solution, we selected three test datasets. The first two datasets, NYT-256-angular and fashion-mnist-784-euclidean, were sourced from the ANN-Benchmarks suite [16]. The third dataset, Pyemb-50GB, contains 28440005 vector embeddings of dimensionality 384. It was compiled during prior research [17] and was specifically chosen to test the proposed solution's ability to handle large volumes of vector embeddings. Compared to the other two datasets, vectors stored in PyEmb-50GB also have a secondary identifier, representing the index of one of ten vector embedding clusters produced based on the contents of that document.

For the comparison, each dataset was stored in the described storage system, having the document index as a primary identifier of the vector embedding. As explained above, the additional metadata contains a secondary identifier, the value of which was taken from the actual secondary identifier for the PyEmb-50GB dataset, and zero for the first two datasets.

For each of the resulting storage instances, the time required to access a single vector using its primary and secondary identifiers was benchmarked.

The proposed solution is compared to general-purpose data storage systems: SQLite3 [18], H2 [19], Apache ORC and Apache Parquet. For all four libraries, a so-called VectorStorage interface was



```
{
  "namespace": "ru.ifmo.ve",
  "type": "record",
  "name": "VectorStorageEntries",
  "fields": [
    {
      "name": "segmentIndexes",
      "type": {
        "type": "array",
        "items": "long"
      }
    },
    {
      "name": "vectors",
      "type": {
        "type": "bytes"
      }
    }
  ]
}
```

Fig. 3. Avro schema of the entries on the page

implemented in Kotlin, along with the fifth implementation that uses the proposed storage approach. Then, these five implementations were benchmarked.

Each test dataset was stored in the specified storage systems using the available compression algorithms with settings providing the best compression, meaning the smallest possible storage size. For SQLite3, individual vectors were compressed, for Parquet and ORC, their own compression was used. For Parquet and ORC, their key-based indexing methods were also enabled to speed up data retrieval by the vector embedding document index. For H2, its built-in database-level compression was used. For Parquet and ORC, a simple entry structure with three fields was used: document index, segment index, and vector embedding, – while the document index was selected as the primary identifier. For SQLite and H2, a table with these three fields was created, and the identifiers were selected as the composite primary key.

The test server has the following specifications: AMD Ryzen 7 7700X (8C16T); 32GB RAM; Operating System: Ubuntu 22.04; OpenJDK 22; a framework of comparing vector search algorithms, implemented in previous research [20], that uses the Java Microbenchmark Harness (JMH).

Experiment results

The first dataset, fashion-mnist-784-euclidean, consists of 50000 sparse vector embeddings with the size of 784, each being a grayscale 28 by 28-pixel image. This dataset compresses well due to the presence of repeated zero components, representing black pixels, in the vector embeddings.

The file sizes of the storage systems and the average retrieval time for a single vector are presented as raw data in Table 1 and visualized in Fig. 4. A dash indicates that the measurement is unavailable because the compression algorithm is not supported for that storage system. For this dataset, the best solution in terms of disk space usage is Parquet storage with the deflate compression algorithm. Parquet occupies the smallest memory volume even without using any compression algorithms due to its built-in RLE mechanism, which performs well on repetitive data, such as sparse vectors.

The proposed solution, with N = 100 and the LZMA compression algorithm, slightly outperforms Parquet in terms of disk space usage (by 0.9 MB, or 3%), but significantly reduces the vector retrieval time by a factor of 100. Increasing the page size with any compression algorithm results in a slight reduction in disk space usage, but retrieval time increases linearly. The retrieval time is comparable to SQLite and H2, except for compressed H2, but they require more disk space than the proposed solution. Compressed H2 demonstrates much slower retrieval time due to the overhead necessary to decompress the entire database.

The second dataset, NYT-256-angular, contains 290000 vector embeddings with the size of 256. Unlike the first dataset, NYT-256-angular consists of dense vector embeddings created from text articles.

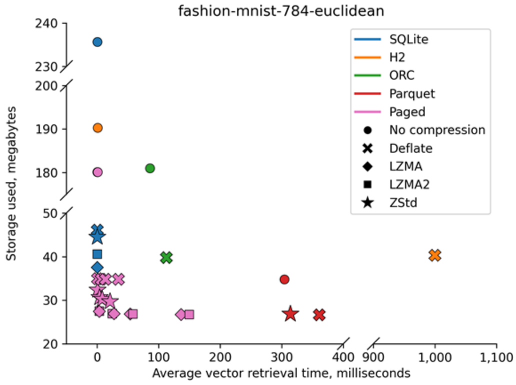


Fig. 4. Disk size (in megabytes) over single vector retrieval time (in milliseconds) for the fashion-mnist-784-euclidean dataset

Table 1

Disk size (in megabytes) and single vector retrieval time (in milliseconds) for the fashion-mnist-784-euclidean dataset

Storage	Storage used, megabytes					Average vector retrieval time, milliseconds				
	w/o	Deflate	LZMA	LZMA2	ZStd	w/o	Deflate	LZMA	LZMA2	ZStd
SQLite	235.7	46.1	37.5	40.6	44.5	0.1 ± 0.0	0.1 ± 0.0	0.3 ± 0.0	0.3 ± 0.0	0.1 ± 0.0
H2	190.3	40.3	–	–	–	0.5 ± 0.0	1000.0 ± 160.9	–	–	–
ORC	181.0	39.8	–	–	–	85.6 ± 5.0	112.3 ± 1.5	–	–	–
Parquet	34.8	26.6	–	–	26.8	304.1 ± 21.0	360.8 ± 27.3	–	–	313.8 ± 22.0
Paged, N = 100	180.1	34.9	27.4	27.5	32.3	0.2 ± 0.0	0.9 ± 0.0	3.5 ± 1.1	3.4 ± 0.1	0.7 ± 0.2
Paged, N = 1000	180.1	34.8	26.9	26.9	30.6	0.2 ± 0.0	6.4 ± 0.2	27.3 ± 1.6	24.5 ± 1.5	4.1 ± 0.1
Paged, N = 2000	180.1	34.8	26.8	26.8	30.2	0.3 ± 0.0	13.3 ± 0.5	53.8 ± 4.3	58.2 ± 4.5	8.3 ± 0.3
Paged, N = 5000	180.1	34.8	26.7	26.7	29.7	0.7 ± 0.0	34.5 ± 1.6	136.2 ± 13.1	149.5 ± 14.3	20.9 ± 1.0

The file sizes of the storage systems and the average retrieval time for a single vector are presented as raw data in Table 2 and visualized in Fig. 5. For this dataset, the proposed solution achieved the smallest storage size among all the available solutions. Using the ZStd compression algorithm and a page size of N = 100, the proposed solution uses 0.3 MB less (1%) than Parquet with the same compression algorithm, while the retrieval time for a vector is reduced by a factor of 163. In the task of compressing dense vector embeddings, the best results were obtained with the ZStd compression algorithm as the page size N increases. However, the retrieval time for a single vector also increases linearly with N. SQLite and H2 still provide comparable retrieval time, except for compressed H2, while requiring significantly more disk space.

The third dataset, PyEmb-50GB, contains 28440005 dense vector embeddings with the size of 384. Unlike previous experiments, for this dataset, larger values of N were used to prevent an increase in the

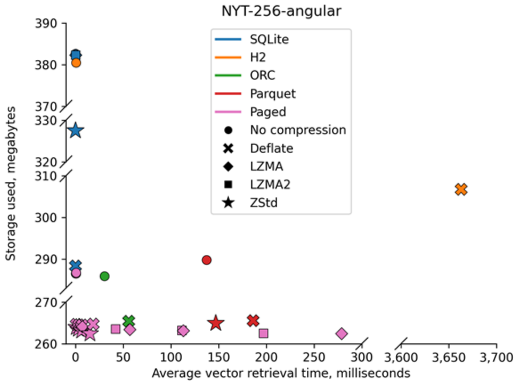


Fig. 5. Disk size (in megabytes) over single vector retrieval time (in milliseconds) for the NYT-256-angular dataset

Table 2

Disk size (in megabytes) and single vector retrieval time (in milliseconds) for the NYT-256-angular dataset

Storage	Storage used, megabytes					Average vector retrieval time, milliseconds				
	w/o	Deflate	LZMA	LZMA2	ZStd	w/o	Deflate	LZMA	LZMA2	ZStd
SQLite	382.6	288.3	382.3	382.3	327.5	0.1 ± 0.0	0.1 ± 0.0	0.3 ± 0.0	0.3 ± 0.0	0.1 ± 0.0
H2	380.5	306.7	–	–	–	0.5 ± 0.0	3663.0 ± 311.8	–	–	–
ORC	285.9	265.4	–	–	–	30.4 ± 0.8	55.9 ± 2.2	–	–	–
Parquet	289.8	265.5	–	–	264.9	137.2 ± 11.6	186.0 ± 17.3	–	–	147.0 ± 9.4
Paged, N = 100	286.5	264.6	264.1	264.2	263.9	0.6 ± 0.1	1.0 ± 0.1	7.2 ± 0.4	4.9 ± 0.2	0.9 ± 0.1
Paged, N = 1000	286.5	264.5	263.4	263.5	263.5	0.2 ± 0.0	3.6 ± 0.1	57.1 ± 4.6	42.1 ± 4.0	3.0 ± 0.1
Paged, N = 2000	286.5	264.5	263.1	263.2	263.1	0.3 ± 0.0	6.9 ± 0.2	113.2 ± 11.3	111.6 ± 12.7	6.3 ± 0.2
Paged, N = 5000	286.7	264.7	262.4	262.5	262.4	0.7 ± 0.0	18.4 ± 0.7	278.8 ± 34.6	197.0 ± 34.3	15.2 ± 0.8

number of files in the storage, which would lead to a significant growth in disk space usage due to the specifics of storing very small files.

The file sizes of the storage systems and the average retrieval time for a single vector are presented as raw data in Table 3 and visualized in Fig. 6. For this dataset, the proposed storage solution uses less disk space for every combination of page size and compression algorithm tested. With N = 10000 and the ZStd compression algorithm, the proposed solution reduces the storage size by 3.7 GB (15%) compared to Parquet with the same compression algorithm, while the retrieval time for a single vector is 14 times faster. Although SQLite Storage still requires the least time to retrieve a single vector, it uses the largest memory volume, and its memory size does not decrease significantly, when compression algorithms are applied, due to inefficiencies of compressing individual vectors. H2 demonstrates better compression due to the database-level compression, but suffers from significantly slower access times.

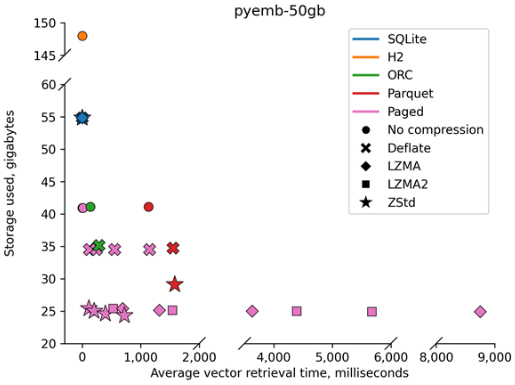


Fig. 6. Disk size (in gigabytes) over single vector retrieval time (in milliseconds) for the PyEmb-50GB dataset

Table 3

Disk size (in gigabytes) and single vector retrieval time (in milliseconds) for the PyEmb-50GB dataset

Storage	Storage used, gigabytes					Average vector retrieval time, milliseconds				
	w/o	Deflate	LZMA	LZMA2	ZStd	w/o	Deflate	LZMA	LZMA2	ZStd
SQLite	54.8	54.8	54.8	54.8	54.8	0.3 ± 0.0	0.2 ± 0.0	0.7 ± 0.0	0.6 ± 0.0	0.3 ± 0.0
H2	148.0	92.5	–	–	–	0.6 ± 0.0	9832.5 ± 761.9	–	–	–
ORC	41.1	35.1	–	–	–	141.0 ± 7.8	286.0 ± 21.3	–	–	–
Parquet	41.1	34.7	–	–	29.1	1132.5 ± 32.5	1553.6 ± 191.1	–	–	1582.5 ± 956.6
Paged, N = 100	40.9	34.5	25.4	25.4	25.4	12.12 ± 2.5	125.4 ± 15.5	690.7 ± 361.1	529.7 ± 223.3	113.3 ± 16.5
Paged, N = 1000	40.9	34.5	25.1	25.1	25.0	11.3 ± 0.7	241.0 ± 7.0	1321.2 ± 816.1	1543.2 ± 782.1	202.1 ± 34.5
Paged, N = 2000	40.9	34.5	25.0	25.0	24.6	3.0 ± 0.5	552.8 ± 108.9	3632.5 ± 2137	4389.0 ± 1725.9	394.7 ± 73.3
Paged, N = 5000	40.9	34.5	24.9	24.9	24.3	4.7 ± 0.4	1149.9 ± 94.0	8754.4 ± 808.1	5674.9 ± 3995.1	722.4 ± 248.0

Conclusion

Using the proposed a page-based storage approach for vector embeddings, combined with the use of general-purpose lossless compression algorithms, reduces the occupied disk space by 14–40% compared to existing solutions for big data storage, such as ORC and Parquet, and up to two times compared to the universal SQLite solution and H2. This was demonstrated in the experiment with the PyEmb-50GB dataset. The solution also reduces the access time by up to a hundred times compared to ORC and Parquet, although it is still slower than SQLite. The ZStd compression algorithm showed good results in experiments with dense vector embeddings, while sparse vector embeddings were more efficiently compressed using the LZMA and LZMA2 algorithms. Increasing the value of N results in a linear increase in access speed to a single vector embedding, while the storage size decreases logarithmically.

The presented storage organization approach can be used in various applications where it is necessary to store vector embeddings, such as information retrieval systems or recommendation systems. Due to its ability to group vector embeddings, it can also be used when implementing average nearest neighbors search systems, which sets it apart from other data formats popular in the industry.

REFERENCES

1. **Grbovic M., Cheng H.** Real-time personalization using embeddings for search ranking at Airbnb. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (*KDD'18*), 2018, Pp. 311–320. DOI: 10.1145/3219819.3219885

2. **Berry M.W., Drmac Z., Jessup E.R.** Matrices, vector spaces, and information retrieval. *SIAM Review*, 1999, Vol. 41, No. 2, Pp. 335–362. DOI: 10.1137/S0036144598347035

3. **Tomilov N.A., Turov V.P., Babayants A.A., Platonov A.V.** A method of storing vector data in compressed form using clustering. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, Vol. 24, No 1, Pp. 112–117. DOI: 10.17586/2226-1494-2024-24-1-112-117

4. **Wu Z.-b., Yu J.-q.** Vector quantization: a review. *Frontiers of Information Technology & Electronic Engineering*, 2019, Vol. 20, Pp. 507–524. DOI: 10.1631/FITEE.1700833

5. **Zhang J., Yang J., Yuen H.** Training with low-precision embedding tables, Available: http://learningsys.org/nips18/assets/papers/78CameraReadySubmissionlp_training_final_v3.pdf (Accessed 30.05.2025)

6. **Zeng X., Hui Y., Shen J., Pavlo A., McKinney W., Zhang H.** An empirical evaluation of columnar storage formats. *Proceedings of the VLDB Endowment*, 2023, Vol. 17, No. 2, Pp. 148–161. DOI: 10.14778/3626292.3626298

7. **Ivanov T., Pergolesi M.** The impact of columnar file formats on SQL-on-hadoop engine performance: A study on ORC and Parquet. *Concurrency and Computation: Practice and Experience*, 2020, Vol. 32, No. 5, Art. no. e5523. DOI: 10.1002/cpe.5523

8. **Agarwal S., Singh N.K., Meel P.** Single-document summarization using sentence embeddings and k-means clustering. *2018 International Conference on Advances in Computing, Communication Control and Networking* (*ICACCCN*), 2018, Pp. 162–165. DOI: 10.1109/ICACCCN.2018.8748762

9. **Dhanabal S., Chandramathi S.** A review of various k-nearest neighbor query processing techniques. *International Journal of Computer Applications*, 2011, Vol. 31, No. 7, Pp. 14–22.

10. **Manro A., Kriti, Sinha S., Chaturvedi B., Mohan J.** Index seek versus table scan performance and implementation of RDBMS. *Advances in Signal Processing and Communication*, 2019, Pp. 411–420. DOI: 10.1007/978-981-13-2553-3_40

11. **Li W., Zhang Y., Sun Y., Wang W., Li M., Zhang W.** Approximate nearest neighbor search on high dimensional data – Experiments, analyses, and improvement. *IEEE Transactions on Knowledge and Data Engineering*, 2019, Vol. 32, No. 8, Pp. 1475–1488. DOI: 10.1109/TKDE.2019.2909204

12. **Golomb S.** Run-length encodings (Correspondence). *IEEE Transactions on Information Theory*, 1966, Vol. 12, No. 3, Pp. 399–401. DOI: 10.1109/TIT.1966.1053907

13. **Deutsch P.** DEFLATE Compressed Data Format Specification version 1.3. *RFC 1951*, 1996. DOI: 10.17487/RFC1951

14. **Berz D., Engstler M., Heindl M., Waibel F.** Comparison of lossless data compression methods. *Technical Reports in Computing Science No. CS-07-2015*, 2015, Vol. 2015, No. 1, Pp. 1–13.

15. **Collet Y., Kucherawy M.** Zstandard compression and the 'application/zstd' media type. *RFC 8878*, 2021, Pp. 1–45. DOI: 10.17487/RFC8878

16. **Aumüller M., Bernhardsson E., Faithfull A.** ANN-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Similarity Search and Applications* (*SISAP 2017*), 2017, Pp. 34–49. DOI: 10.1007/978-3-319-68474-1_3

17. **Tomilov N., Turov V., Babayants A.** Algoritmy vektornogo poiska v zadache tekstovogo poiska [Vector search algorithms in text search]. *XII Kongress molodykh uchenykh* [*XII Congress of Young Scientists*], 2023, Vol. 1, Pp. 406–411.

18. **Bi C.** Research and application of SQLite embedded database technology. *WSEAS Transactions on Computers*, 2009, Vol. 8, No. 1, Pp. 83–92.

19. **Diniz Junior R.N.V., da Rocha R.F., dos Santos L.M., Junior M.R.G.B., Costa Bezerra E.** A comparison of in-memory databases in Java application. *2024 IEEE 4th International Conference on Information Technology, Big Data and Artificial Intelligence* (*ICIBA*), 2024, Vol. 4, Pp. 665–670. DOI: 10.1109/ICIBA62489.2024.10868437

20. **Tomilov N., Turov V., Babayants A.** Razrabotka instrumenta sravneniia algoritmov vektornogo poiska [Developing a comparison tool for vector search algorithms]. *XII Kongress molodykh uchenykh* [*XI Congress of Young Scientists*], 2022, Vol. 1, Pp. 446–450.

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Tomilov Nikita A.
Томилов Никита Андреевич
E-mail: programmer174@icloud.com
ORCID: https://orcid.org/0000-0001-9325-0356

Turov Vladimir P.
Туров Владимир Павлович
E-mail: firemoon@icloud.com
ORCID: https://orcid.org/0009-0009-1470-7633

Submitted: 06.01.2025; Approved: 26.05.2025; Accepted: 30.05.2025.
Поступила: 06.01.2025; Одобрена: 26.05.2025; Принята: 30.05.2025.

Research article
DOI: https://doi.org/10.18721/JCSTCS.18205
UDC 004.03

CONCEPT OF ENSURING THE RESILIENCE OF OPERATION OF NATIONAL DIGITAL PLATFORMS AND BLOCKCHAIN ECOSYSTEMS UNDER THE NEW QUANTUM THREAT TO SECURITY

V.Yu. Skiba[1] ✉, *S.A. Petrenko*[1], *K.O. Gnidko*[1], *A.S. Petrenko*[2]

[1] Sirius University of Science and Technology, Federal Territory “Sirius”,
Krasnodar Krai, Russian Federation;

[2] St. Petersburg Electrotechnical University,
St. Petersburg, Russian Federation

✉ vskiba69@mail.ru

Abstract. The obtained results in the field of quantum informatics clearly demonstrate the high technological potential of quantum technologies. A cryptanalytically relevant or significant quantum computer can threaten the operation of various systems, including national blockchain ecosystems and platforms in the Russian Federation. In this situation, a concept of ensuring the resilience of the operation of national digital platforms and blockchain ecosystems under the new quantum security threat is needed, the provisions of which are substantiated in this article. The concept contains a justification for the relevance of the problem and strategic goals of ensuring quantum resilience, national interests in the field of quantum information technologies, the presence of quantum threats to the operation of digital platforms and blockchain ecosystems, methods, means and priority measures to ensure the quantum resilience of national digital platforms and blockchain ecosystems. The main directions of further research of the group “Technologies for countering previously unknown quantum cyber threats” of the Scientific Center for Information Technology and Artificial Intelligence of the Sirius University of Science and Technology, aimed at implementing the proposed concept, are also considered.

Keywords: information security, quantum information technology, quantum security threat, quantum resilience, technological safety, operational safety, verifiable safety

Acknowledgements: The project “Technologies for countering previously unknown quantum cyber threats” was selected for support within the framework of event 2.3 of the state program of the federal territory Sirius “Scientific and technological development of the federal territory Sirius” (application registration No. FCS-2024-2.3-VY-1160-5744, leading scientist – Petrenko S.A., PhD).

Citation: Skiba V.Yu., Petrenko S.A., Gnidko K.O., Petrenko A.S. Concept of ensuring the resilience of operation of national digital platforms and blockchain ecosystems under the new quantum threat to security. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 56–73. DOI: 10.18721/JCSTCS.18205

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18205
УДК 004.03

КОНЦЕПЦИЯ ОБЕСПЕЧЕНИЯ УСТОЙЧИВОСТИ ФУНКЦИОНИРОВАНИЯ НАЦИОНАЛЬНЫХ ЦИФРОВЫХ ПЛАТФОРМ И БЛОКЧЕЙН-ЭКОСИСТЕМ В УСЛОВИЯХ НОВОЙ КВАНТОВОЙ УГРОЗЫ БЕЗОПАСНОСТИ

В.Ю. Скиба[1] ✉, *С.А. Петренко*[1], *К.О. Гнидко*[1], *А.С. Петренко*[2]

[1] Научно-технологический университет «Сириус», федеральная территория «Сириус», Краснодарский край, Российская Федерация;

[2] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» имени В.И. Ульянова (Ленина), Санкт-Петербург, Российская Федерация

✉ vskiba69@mail.ru

Аннотация. Полученные результаты в области квантовой информатики наглядно показывают высокий технологический потенциал квантовых технологий. Криптоаналитически релевантный или значимый квантовый компьютер может поставить под угрозу функционирование различных систем, в том числе национальных блокчейн-экосистем и платформ в Российской Федерации. В этой ситуации необходима концепция обеспечения устойчивости функционирования национальных цифровых платформ и блокчейн-экосистем в условиях новой квантовой угрозы безопасности, положения которой обоснованы в данной статье. Концепция содержит обоснование актуальности проблемы и стратегических целей обеспечения квантовой устойчивости, национальных интересов в сфере квантовых информационных технологий, наличия квантовых угроз для функционирования цифровых платформ и блокчейн-экосистем, методов, средств и первоочередных мероприятий обеспечения квантовой устойчивости национальных цифровых платформ и блокчейн-экосистем. Также рассматриваются основные направления дальнейших исследований группы «Технологии противодействия ранее неизвестным квантовым киберугрозам» НЦ ИТ и ИИ Научно-технологического университета «Сириус», направленные на реализацию предложенной концепции.

Ключевые слова: информационная безопасность, квантовая информационная технология, квантовая угроза безопасности, квантовая устойчивость, технологическая безопасность, эксплуатационная безопасность, верифицируемая безопасность

Финансирование: Проект «Технологии противодействия ранее неизвестным квантовым киберугрозам» был отобран для поддержки в рамках мероприятия 2.3 государственной программы федеральной территории «Сириус» «Научно-технологическое развитие федеральной территории "Сириус"» (регистрационный номер заявки ФТС-2024-2.3-VY-1160-5744, ведущий ученый – д.т.н. Петренко С.А.).

Для цитирования: Skiba V.Yu., Petrenko S.A., Gnidko K.O., Petrenko A.S. Concept of ensuring the resilience of operation of national digital platforms and blockchain ecosystems under the new quantum threat to security // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 56–73. DOI: 10.18721/JCSTCS.18205

Introduction

With the advent of the first cryptocurrency platform Bitcoin in 2009, the implementation of blockchain technology began as a special case of building distributed ledger systems [1, 2]. With the advent of the Ethereum platform, it became possible to save records in a ledger created by the user, develop smart contracts to describe the business logic of any transactions without intermediaries [2]. These capabilities made it possible to store all the necessary information (from the conclusion of a contract to the successful closure of a transaction, subsequent warranty service and compliance with the rights to the intellectual property used) in distributed ledgers in the format of smart contracts, and were developed numerous concepts for creating digital information and logistics platforms to control all regulatory operations [3, 4]. As noted in [3, 4], the implementation of such digital information and logistics platforms allows reducing the risks of logistics errors, unifying document flow, visualizing in real time the processes of movement of goods and the status of customs documents.

All this has led to the rapid development of blockchain technologies (Enterprise Ethereum Alliance, Waves Enterprise, Hyperledger Fabric, Corda Enterprise, Bitfury Exonum, Blockchain Industrial Alliance, Exonum, Nodes Plus, Microsoft Azure Blockchain, Masterchain 2.0 etc.) both in terms of the emergence of new blockchain platforms for various purposes, and in terms of the development of the decentralized financial sector, which de facto is a system of various protocols and applications built on blockchain technology and operating on the basis of smart contracts.

There is a steady growth in the implementation of blockchain technologies in the creation of digital platforms in various areas and industries of the Russian Federation. The National Technology Initiative website[1] lists a number of developed and implemented projects, such as Cryptoveche (remote voting system), Edemes (database for identifying and monitoring the movement of cultural property), InsurDoc (management of intellectual property rights), Trevo (control of cargo and goods transportation), DSMS (decentralized exchange of financial messages) and a number of other projects. Web3 Tech company offers a range of products operating on its own blockchain platform called Confident. Among the products offered, based on the Confident platform, is the software and hardware complex for working with digital assets, Digital Treasury.

Using similar approaches, the Bank of Russia "Digital Ruble" platform and the Russian Export Center's information system "My Export" digital platform (state information system "Single Window") are evolving, and the National digital transport and logistics platform (NDTLP) is being created [4, 5].

Simultaneously with this, the issue of the security of using blockchain technologies is constantly being raised. First of all, these issues are related to the security and reliability of storing and using data placed in systems using blockchain technologies (blockchain ecosystems). An analysis of publications shows that researchers, as a rule, focus their attention on one or more interrelated aspects that pose a potential threat to the operation of the blockchain ecosystem.

The development of quantum computing poses a threat to existing cryptographic mechanisms used in digital platforms and blockchain ecosystems (DPiBE), creating risks to transaction security and threatening the integrity and immutability of data in distributed ledger systems.

The results obtained in the field of quantum technologies and quantum information technologies clearly demonstrate the high technological potential of quantum technologies for solving a number of problems, much more efficiently than any modern "traditional" computer [2, 4–7].

A cryptoanalytically relevant or significant quantum computer may threaten the resilience of various systems [4, 8, 9], including critical information infrastructure facilities of the Russian Federation, national DPiBE.

Ensuring the resilience of DPiBE to attacks by intruders (or malicious actors) using a quantum computer (or, in other words, their quantum resilience) is one of the pressing scientific and technical problems of the digital economy of the Russian Federation.

In this situation, the Concept is needed to ensure the resilience of the operation of national DPiBE under of a new threat to quantum security (the Concept), which should contain, according to the opinion of the authors of this article, at least the following provisions based on the systematization of the results obtained in more than 30 other our own papers:

- national interests and strategic goals of the Russian Federation;

[1] Natsional'naia tekhnoløgicheskaia initsiativa (NTI) [National Technology Initiative (NTI)], Available: https://nti2035.ru/ (Accessed 13.03.2025)

- current state of development of quantum technologies and quantum information technologies;
- main quantum threats to DPiBE and the ways of ensuring their quantum resilience.

Prerequisites for conducting research in the sphere of quantum resilience of national DPiBE

Despite numerous doubts about the feasibility of using blockchain technologies in the early years of their development, today there is a stable growth and development of these technologies, as well as their implementation in various digital platforms in almost all areas and sectors of the digital economy of the Russian Federation. At the same time, questions arise about ensuring the security of information in DPiBE.

The global information space, DPiBE are simultaneously used, on the one hand, to expand access of individuals and legal entities to information, digital services and financial instruments and to increase the efficiency of the digital economy (data economy), and, on the other hand, by criminal structures, international terrorist organizations and states pursuing an unfriendly policy towards the Russian Federation, to disrupt the functionality of these facilities and create centers of social tension [4, 8–12].

Conducting cyber operations against transportation infrastructure, power grids, dams, chemical plants, nuclear power plants and other critical infrastructure, DPiBE is technically possible. Such operations could have large-scale consequences, causing significant damage and/or leading to a large number of casualties among the civilian population [4, 8–10, 12].

As DPiBE develop, they acquire new and increasingly emergent system properties: controllability, self-organization, adaptability, cybersecurity, technological security, operational security, verifiable security and cyber resilience (including quantum resilience). Each of these properties is the subject of research, and each subsequent property makes sense only if the previous one is present.

Cyber resilience of a DPiBE is understood as the ability of a DPiBE, operating according to a certain set of algorithms, to achieve the goals and objectives of operation in the face of growing security threats.

Summarizing the results of the authors' research (for example, [4, 8, 9, 13]), the following conclusions can currently be drawn:

- in the context of the growth of classical and quantum cyberattacks by intruders, ensuring cyber resilience is becoming much more difficult;
- of particular concern are the so-called new type of quantum threats – quantum attacks or attacks using a quantum computer;
- most cryptographic primitives used in modern information systems (including hash functions, digital signatures, asymmetric cryptographic algorithms and related protocols) are not resistant to quantum attacks;
- in order to hack the crypto-primitives used, a number of well-known quantum algorithms can be successfully and effectively applied, in particular, Shor's algorithm [14, 15] for factorization and discrete logarithm and Grover's algorithm for accelerating the attack on the hash function [16].

This creates risks to transaction security and threatens the integrity and immutability of data in distributed ledger systems.

Accordingly, the quantum resilience of DPiBE is the ability of these systems to achieve the goal of functioning under the conditions of attacks by intruders using a quantum computer.

A bibliometric analysis of scientific literature on quantum technologies for the period from 1990 to 2020 showed [7]: the dynamism of the development of the field, a high degree of concentration of research and international scientific relations, as well as the participation of not only universities and academic organizations, but also large corporations (especially from Japan) and military research structures (primarily from the USA). At the same time, the Russian Federation is characterized by:

- high concentration of research in metropolitan areas and their significant internationalization;
- leading contribution of the Russian Academy of Sciences (RAS), which ranks sixth among scientific organizations in the world in the number of publications in the field of quantum technologies for the analyzed period [6]);

- growing role of universities in the development of the scientific base of quantum technologies and quantum information technologies;
- still weak involvement of the Russian commercial sector in research.

Since 2020, the number of scientific publications on the results of research in the field of quantum technologies and quantum information technologies has been growing exponentially, including due to work on ensuring information security in connection with the emergence of new quantum threats.

The situation in the field of quantum computing is characterized by a kind of "technological race" between leading companies [4]. There are dozens of organizations in the world attracting significant investments to create quantum computers[2].

In the Russian Federation, scientific research and engineering surveys are also being conducted to create the first domestic quantum computers. Well-known Russian mathematicians have made significant contributions to this field of knowledge, for example, employees of the Steklov Mathematical Institute of RAS: head of the Department of Mathematical Physics, PhD, corresponding member of the RAS I.V. Volovich and head of the Department of Probability Theory and Mathematical Statistics, laureate of the Claude E. Shannon Award for outstanding achievements in information theory, PhD, academician of the RAS A.S. Holevo [17, 18]. And scientists from the Russian Quantum Center and P.N. Lebedev Physical Institute of RAS, for example, have developed a prototype of a quantum computer on ytterbium ions[3]. Quantum processors with 2–10 qubits and quantum simulators with 10–20 qubits have been developed, and the first domestic quantum processors with 50–100 qubits are expected to appear by the end of 2025.

Despite the fact that the era of noisy intermediate-scale quantum (NISQ) devices is currently ongoing, quantum information science as a whole is already a new, rapidly developing branch of science associated with the use of quantum systems to implement fundamentally new methods of transmitting messages, computing and technologies (quantum communication channels, quantum cryptography, quantum computer) [9, 17–21].

Periodically, there are reports of achieving "quantum supremacy"[4], that is, the creation of a quantum computer capable of solving problems significantly more efficiently than any modern "traditional" computer (modern von Neumann supercomputers are also considered "traditional" tools in this approach) or even impossible to solve using "traditional" computing tools [4, 22]. Soon, quantum computers will reach sufficient maturity and will be able to "hack" most cryptographic primitives used in blockchain ecosystems [4, 23, 24].

These results have allowed experts to predict that quantum information technologies capable of cracking Bitcoin cryptographic algorithms could be created in 2027, and the RSA cryptographic algorithm in 2031 [25]. The British regulator (National Cyber Security Centre, NCSC) in its 2020 recommendations predicts the emergence of a cryptographically significant quantum computer in 2030[5].

It should be taken into account that most forecasts are based on open data, and in the conditions of geopolitical confrontation, there is a possibility that real successes in creating a working quantum computer are confidential information. It is possible to determine the real situation only after identifying facts of compromise of a significant array of data or facts of disruption of the functioning of any systems (not necessarily DPiBE) as a result of the use of quantum computers by an intruder, including in conjunction with "traditional" computers.

[2] Quantum computing start-up secures €10m investment – National Technology, Available: nationaltechnology.co.uk (Accessed 13.03.2025); PsiQuantum Closes $450 Million Funding Round to Build the World's First Commercially Viable Quantum Computer &mdash; PsiQuantum, Available: https://www.psiquantum.com/news-import/psiquantum-closes-450-million-funding-round-to-build-the-worlds-first-commercially-viable-quantum-computer (Accessed 13.03.2025)

[3] Sozdan prototip kvantovogo komp'iutera na ionakh itterbia [A prototype of a quantum computer on ytterbium ions has been created], Available: https://strana-rosatom.ru/2022/02/25/sozdan-prototip-kvantovogo-kompjute/ (Accessed 13.03.2025)

[4] Here the question of "price vs quality" arises: how much more expensive is such a quantum component of a computing system than a traditional one, capable of doing the same work, albeit over a longer period of time?

[5] Preparing for Quantum-Safe Cryptography, Available: https://www.ncsc.gov.uk/whitepaper/preparing-for-quantum-safe-cryptography (Accessed 13.03.2025)

Thus, developments in the field of creating quantum computers and developing quantum information technologies with quantum computing algorithms create more and more preconditions for their use by potential intruders (or malicious actors) to disrupt the operation of DPiBE.

Scenario and methods for ensuring quantum resilience of DPiBE

In [4, 8], it has already been noted that a number of technological countries around the world have already begun to prepare to counteract future quantum threats.

The US presidential administration has issued several directives[6] on preparing the state and business for future quantum cyberattacks, and has also instructed the National Institute of Standards and Technology (NIST), the National Security Agency (NSA), and the Cybersecurity and Infrastructure Security Agency (CISA) to take all necessary measures to protect the critical infrastructure of the US and its NATO allies from a quantum threat within one year. There is no information on what has been done.

In [4, 13], it is substantiated that due to the current lack of a unified scientific, methodological and technical basis for the development of quantum resilience DPiBE in the Russian Federation, the creation of such systems is a long-term task. At the same time, it is necessary to develop a new technology to counter quantum security threats. Without solving this fundamental scientific problem, it is impossible to talk about achieving the goals of the national program "Data Economy"[7].

In this regard, in the fall of 2024, a research group "Technologies for countering previously unknown quantum cyber threats" was formed at the Scientific Center for Information Technology and Artificial Intelligence of the Sirius University of Science and Technology.

The main goal of creating the research group is to create a promising world-class technology to ensure quantum resilience of the leading national DPiBE of the digital economy of the Russian Federation, which, unlike known technologies, will prevent significant or catastrophic consequences in the face of previously unknown cyberattacks by intruders using a quantum computer.

Based on the results obtained in [4, 23, 24], when developing the Concept, it is necessary to take into account two main scenarios for the development of quantum information technologies:

1. Use of quantum computing tools in the infrastructure of DPiBE for information processing and/or protection. Obviously, in this case, quantum computing tools and quantum calculations will be used together with "traditional" computing tools and information protection;

2. Use of quantum computing tools to solve individual local computationally intensive problems. At the same time, the formation of "hostile" information impacts is realized using quantum computing tools.

The absence of serial production of basic elements of quantum computing tools ("quantum chips"), as well as the selected basic physical platform for the creation and production, at least in small batches, of quantum processors (or computers) determines the second scenario as a priority.

The results of an analysis of various platforms (using which work is being carried out to create quantum computers, emulate quantum computing, and develop quantum information processing technologies) and quantum threats to various DPiBE (for example, performed in [2, 4, 8, 9]) allow us to draw a number of conclusions:

- development of quantum computing threatens the existing cryptographic mechanisms used in DPiBE;
- achievements of IBM, as well as a number of other high-tech manufacturers of quantum computers, convincingly demonstrate the realism of the implementation of quantum threats in the near future. The emergence of a relevant quantum computer capable of cracking traditional cryptographic algorithms is expected in the period 2026–2030;

[6] Memorandum on Preparing for Post-Quantum Cryptography, Available: https://www.dhs.gov/publication/memorandum-preparing-post-quantum-cryptography (Accessed 13.03.2025); National Security Memorandum on Promoting United States Leadership in Quantum Computing While Mitigating Risks to Vulnerable Cryptographic Systems, Available: Preparing Secrets for a Post-Quantum World—National Security Memorandum 10 – EveryCRSReport.com (Accessed 13.03.2025)

[7] The national program "Data Economy" has been implemented in the Russian Federation since 2025 and replaces the project "Digital Economy of the Russian Federation", which was completed in 2024.

- main problematic issues of ensuring quantum resilience include the insufficient level of readiness for the growth of quantum cyberattacks by intruders and the growth in the number and complexity of DPiBE structures, as well as the difficulty of identifying quantitative patterns that allow us to study the cyber stability of DPiBE in the face of classical and quantum cyberattacks by malicious actors.

Ignorance or ignoring the above-mentioned problematic issues leads to a decrease in the efficiency of DPiBE.

The Concept also needs to take into account that the use of classical approaches (for example, methods of mathematical statistics and experimental design and analytical verification methods) considered in [4] to identify the indicated patterns is impractical due to the presence of contradictions associated with the specifics of specific models of system behavior in the presence of quantum and classical attacks by an intruder.

The Concept must provide for the use of methods for ensuring verified security, when technological and operational security is ensured using one mathematical apparatus of specification and verification both at the stage of system creation and adaptive information security management using a reference model with anticipatory forecasting. Such methods were proposed in [26], but require development in the interests of ensuring quantum resilience of national DPiBE.

It seems advisable to identify three main directions in the Concept for solving the scientific problem of ensuring the quantum resilience of national DPiBE, which, in accordance with [4, 8, 9, 23, 24], have already been developed to one degree or another.

The research group for countering previously unknown quantum cyber threats has completed the development of three new post-quantum algorithms for electronic digital signatures based on the mathematical apparatus of finite non-commutative associative algebras (FNAA). Effective algorithms for solving this problem on classical and quantum computers are currently unknown [27].

Similar works are also known in this area, aimed at creating:

- the post-quantum electronic signature “Rosehip”[8] [28], the cryptographic resistance of which is based on the computationally complex mathematical problem of decoding a random linear code;
- the “Forsythia” protocol for generating a common key using the supersingular elliptic curve isogeny apparatus[9] [29].

This direction is relatively “young” and poorly studied, which requires work on optimizing the performance of post-quantum algorithms and proving their cryptographic resistance in the conditions of using of quantum computers by an intruder. In the spring of 2023, for example, the Royal Swedish Institute of Technology discovered a vulnerability in the CRYSTALS-Kyber post-quantum algorithm, one of the finalists of the famous NIST competition [30].

The use of quantum cryptographic algorithms with mathematically proven cryptographic resistance is the second direction of ensuring quantum resilience of the national DPiBE. This direction should also include the use of quantum data transfer protocols, quantum key distribution protocols, quantum random number generators etc. At the same time, it is necessary to take into account the high probability of the presence of undeclared capabilities and software backdoors in a large number of emerging open and commercial libraries for developers of digital platforms (SDK) implementing new cryptographic schemes.

The third direction of ensuring quantum resilience of the national DPiBE is the creation of a full-fledged quantum infrastructure, which provides for the software and hardware implementation of a fully quantum model of protected information systems. It should be emphasized that this direction is a distant and possibly unachievable prospect in general.

[8] The package of documents on the post-quantum electronic signature “Rosehip” was presented to the Technical Committee for Standardization "Cryptographic Protection of Information" of Rosstandart (TC26) in June 2022.

[9] Developed within the framework of the activities of the working subgroup on isogenies of supersingular elliptic curves of Working Group 2.5 “Post-quantum cryptographic mechanisms” of TC26.

The following section of this article sets out the main provisions of the Concept developed by members of the above-mentioned research group at the first stage of the project by generalizing, systematizing and comprehensively rethinking the results obtained earlier. Given that this section is essentially a draft regulatory legal act, references to the sources used in its writing are not provided.

Proposals for the main provisions of the Concept

The Concept, based on the analysis of the current state of ensuring information security of national DPiBE of the Russian Federation and the development of quantum technologies and quantum information technologies, defines the goals, objectives and key problems of ensuring quantum stability of national DPiBE.

The purpose of the Concept development is to identify the main approaches to achieving the goal of functioning DPiBE of the Russian Federation in the face of attacks by intruders (or malicious actors) using a quantum computer (that is, the quantum resilience of national DPiBE) to ensure global technological competitiveness and technological sovereignty of the Russian Federation in the field of quantum technologies, quantum communications, quantum computing and quantum information technologies.

The Concept should become an integral part of the Concept for Ensuring Information Security of the Russian Federation.

General provisions

The Concept serves as a methodological basis for the development of a set of regulatory and organizational and methodological documents regulating activities in the field of ensuring the quantum resilience of the national DPiBE, as well as for developing proposals to improve scientific, technical and organizational support for the quantum resilience of the national DPiBE, as well as training personnel in this area.

For the purposes of this Concept, the following concepts are used:

- blockchain (or distributed ledger) – a continuous sequential chain of blocks (linked list) built according to certain rules, containing some information;
- blockchain ecosystem – a network of all participants in a blockchain network, who share business processes and business goals;
- quantum computing – a type of computing that uses quantum mechanical phenomena, such as superposition and entanglement to perform operations on data;
- quantum information technology – a technology that provides for the implementation of quantum algorithms using quantum computing, quantum communications or other quantum technologies;
- quantum communication – technology of encoding and transmitting data in quantum states of photons;
- quantum key distribution – procedure of generating and distributing secret keys, implemented using quantum cryptographic protocols and quantum communication channels;
- quantum technology – technology for creating computing systems based on new principles (quantum effects) that allows us to radically change the ways, in which large amounts of information are transmitted and processed;
- quantum resilience – ability to achieve the goal of operation under of attacks by intruders (or malicious actors) using a quantum computer;
- digital platform – online services and/or software products that allow user interaction and transactions, provide access to content, services, or products, and provide user-friendliness for various functions.

Importance of ensuring the quantum resilience of the national DPiBE

Ensuring the resilience and security of critical infrastructure facilities and services for the transmission, processing and storage of large amounts of data in the face of the growth of both classical and previously unknown (and, consequently, poorly studied) security threats is one of the five key objectives of the “Information Security” project in accordance with the requirements of the passport of the national program "Data Economy". Special attention is paid to assessing the level of security of government information systems (GIS).

National DPiBE are generally GIS and are becoming an increasingly important element of the digital economy of the Russian Federation.

The urgency of the problem of ensuring the quantum resilience in fractionally of national DPiBE as GIS is due to:

- the contradiction between the emergence of a new type of information security threat (quantum threats) and the inability of known technologies (models, methods and means) for ensuring information security and cyber resilience to detect, neutralize and prevent such threats;
- the need to increase requirements for ensuring information security and cyber resilience of the critical information infrastructure of the Russian Federation, including national DPiBE;
- the growth in the number of national DPiBE, the increasing complexity of their structures and functioning processes.

National interests in the sphere of quantum technologies and quantum information technologies

The development of the field of quantum technologies and quantum information technologies is one of the tasks aimed at achieving the goal of scientific and technological development of the Russian Federation in accordance with the National Security Strategy[10].

National interests in the field of quantum technologies and quantum information technologies are reflected in the following documents:

- Strategy for scientific and technological development of the Russian Federation[11] (subparagraph "d" of paragraph 21 and subparagraphs "a" – "c" of paragraph 24);
- Priority areas of scientific and technological development (paragraph 4) and the List of the most important science-intensive technologies (paragraph 12)[12];
- List of instructions following the meeting with scientists and the plenary session of the Forum of Future Technologies "Computing and Communications. Quantum World" (regarding the development of quantum technologies and the creation of a university in the field of quantum technologies for the purpose of implementing educational programs for studying advanced developments in this area and involving the participation of schoolchildren)[13];
- Concept for regulating the quantum communications industry in the Russian Federation until 2030[14].

On February 6, 2024, the Federation Council Committee on Defense and Security reviewed and took control of the issue of ensuring information security using quantum technologies as part of the national project to form a data economy.

Strategic goals of ensuring quantum resilience of national DPiBE

The strategic goals of ensuring the quantum resilience of national DPiBE are as follows:

- *Ensuring the quantum sovereignty of the Russian Federation*. The development of quantum technologies and quantum information technologies in the world increases the number of new external risks and threats to the country's digital sovereignty. The emergence of quantum computers, as well as the high degree of development of quantum computing and AI technologies, create risks and threats of compromising existing methods of data protection. It is necessary to ensure the creation

[10] Approved by the decree of the President of the Russian Federation dated July 2, 2021 No. 400 "On the National Security Strategy of the Russian Federation" (http://publication.pravo.gov.ru/document/0001202107030001)

[11] Approved by the decree of the President of the Russian Federation dated February 28, 2024 No. 145 "On the Strategy for Scientific and Technological Development of the Russian Federation" (http://publication.pravo.gov.ru/document/0001202402280003)

[12] Approved by Decree of the President of the Russian Federation dated June 18, 2024 No. 529 "On approval of priority areas of scientific and technological development and a list of the most important science-intensive technologies" (http://publication.pravo.gov.ru/document/0001202406180018)

[13] Approved by the President of the Russian Federation on September 3, 2023 No. Pr-1734 (https://digitalcryptography.ru/news/novosti-otrasli/vladimir-putin-dal-porucheniya-po-razvitiyu-kvantovykh-tekhnologiy/)

[14] Approved by the Order of the Government of the Russian Federation dated July 11, 2023 No. 1856-r "On approval of the Concept for regulating the quantum communications industry in the Russian Federation until 2030" (http://publication.pravo.gov.ru/document/0001202307170029)

and development of domestic quantum information technologies that ensure the quantum resilience of national DPiBE.

- *Creating modern and effective domestic systems for protecting information from quantum threats*. Government agencies, industrial enterprises, scientific and expert communities are working in many directions to create systems capable of countering new quantum threats to information security. It is necessary to ensure the development of a complex of domestic trusted computing equipment using quantum technologies and quantum information technologies and certified information security tools, including cryptographic ones, that ensure effective counteraction to quantum threats.
- *Implementing and integrating quantum and post-quantum cryptography, including methods of quantum communication and quantum key distribution, into key DPiBE in various areas of the digital economy of the Russian Federation*.
- *Providing guarantees to Russian individuals and legal entities to ensure the security of their information processed on national DPiBE*. It is necessary to ensure the attractiveness and competitiveness of using national DPiBE in comparison with foreign DPiBE.
- *Developing the export potential of national DPiBE*. It is necessary to provide proof of the quantum resilience of national DPiBE, recognized by the international community, which will expand the customer base or the number of implementations among states friendly to the Russian Federation.

The current state of ensuring quantum resilience of national DPiBE

In the era of NISQ devices, the components of a quantum computer that can be implemented in practice are imperfect in terms of accuracy and highly susceptible to interference and errors. However, using these components in combination with classical computers and fifth-generation supercomputers will soon allow a malicious actor to achieve significant overall computing acceleration when solving multidimensional optimization and information security problems.

The general situation is that there is no technology for mass production of quantum chips and even a physical platform for quantum computers has not been selected. In parallel, work is underway to create quantum computers based on more than 10 platforms, the main ones being: superconductors, ions, neutral atoms and photons.

The first (on superconductors) are developed by IBM, Google, Rigetti, Intel, Alibaba. The advantages of these platforms include: good scalability, stability over time and relative ease of management. The disadvantages are: need to use ultra-low temperatures and low coherence.

The second (on ions) are being improved by Honeywell, IonQ, AQT. These platforms are characterized by better stability and accuracy of operations. The disadvantage is considered to be the technological limitation of the maximum size of the quantum register.

The third (on neutral atoms) are being improved by Pasqal, Harvard University and the University of Paris-Saclay. Platforms of this type allow for good scaling. At the same time, they are distinguished by the high complexity of managing quantum systems.

The fourth (on photons) are created by Xanadu, Quix, Psi Quantum etc. These platforms are compact in size, operate at room temperatures and are relatively easy to interface with fiber-optic communication lines. However, it is more difficult to implement logical circuits in such platforms due to the weak interaction of photons.

In the Russian Federation, scientific research and engineering studies are also being conducted to create the first domestic quantum computers.

Cooperation and consortia are being formed on the basis of domestic competence centers in the field of quantum technologies, quantum information technologies and quantum communications[15].

[15] The cooperation includes scientific divisions of the Russian Academy of Sciences (Institute of Automation and Electronics SB RAS, SB RAS, IPF RAS and its branch IMF RAS), the Center for Quantum Technologies of Lomonosov Moscow State University, Sirius University of Science and Technology, Bauman Moscow State Technical University, MIET, MIAN., FIAN, PTIAN, ISAN, Russian Quantum Center, FSUE VNIIA named after N.L. Dukhov, ITF named after L.D. Landau, ITF named after A.V. Rzhanov, ISP named after P.L. Kapitsy, ISSP, KNRTU-KAI, KHFTI, KFU, Moscow State Pedagogical University, MIPT, MISiS, NSTU, Skoltech, ITMO University, Ioffe Institute of Physics and Technology and others.

By the end of 2025, the first domestic quantum processors with 50–100 qubits are expected to appear.

At the same time, the results of the analysis of the current state of information security in national DPiBE are showing that the level of quantum resilience currently does not meet the vital needs of individuals, society and the state.

The current conditions of the country's political and social and economic development are exacerbating contradictions between the needs of society to expand the free exchange of information and the introduction of digital, including financial, tools and the need to maintain certain restrictions on the dissemination of information and ensure the sustainability of digital tools.

The main factors of the presence of quantum threats to national DPiBE are:

- insufficient cybersecurity and cyber resilience of DPiBE in the context of the growth of classical and quantum cyberattacks by intruders;
- the growing number of structures of national digital platforms and the growing complexity of the behavior of blockchain ecosystems;
- difficulty in identifying quantitative patterns that allow to study the cyber resilience of national DPiBE in the context of classical and quantum cyberattacks by intruders;
- inconsistency of the actual parameters of the operating of national DPiBE in functional specifications;
- overvaluation of the capabilities of modern methods and means of information protection, reliability and fault tolerance of blockchain.

The lack of effective mechanisms for regulating the quantum resilience of national DPiBE leads to many negative consequences.

Insufficient security of GIS, which are digital platforms or blockchain ecosystems, leads to the loss of important political, scientific, technical, economic or commercial information, including information on foreign economic activity, transport and logistics or other activities important for ensuring the security of the state.

The lack of protection of citizens' rights to information, manipulation of information in GIS, cause an inadequate response from the population and in some cases can lead to political or social instability in society or the state.

The lag of domestic information technologies (including quantum technologies and quantum information technologies) forces operators of national DPiBE to purchase untrusted and unprotected imported computing equipment and software, including blockchain technologies. As a result, the likelihood of unauthorized access to databases and data banks increases, both as a result of classical attacks and especially using quantum threats. The country's dependence on foreign manufacturers of computer equipment, software and information products also increases.

This state of affairs in the field of ensuring quantum resilience of national DPiBE requires solving the following key tasks:

1. Development of scientific and practical foundations of quantum technologies, quantum computing, quantum information technologies and ensuring quantum resilience, corresponding to the world's advanced levels of scientific and technological development, the current geopolitical situation and the conditions of political and social and economic development of the Russian Federation.
2. Improvement of the legislative and regulatory framework for ensuring information security in terms of ensuring cybersecurity and quantum resilience of national DPiBE.
3. Development of modern methods and software and hardware that provide a comprehensive solution to the problems of quantum resilience of national DPiBE.
4. Development of criteria and methods for assessing the quantum resilience of national DPiBE, as well as assessing the effectiveness of systems and means of ensuring the security of national DPiBE.
5. Development of a set of interconnected training programs in the field of quantum information technologies and ensuring quantum resilience of national DPiBE, including additional professional training and/or advanced training.

The quantum threats to the functioning (threats to quantum resilience) of national DPiBE

Sources of threats to the quantum resilience of national DPiBE can be divided into external and internal. The external sources include:

- unfriendly policies of foreign states in the field of global information monitoring, dissemination of information and new information technologies, including quantum technologies and quantum information technologies;
- activities of foreign intelligence services, special services, political and economic structures directed against the interests of the Russian Federation using quantum technologies and quantum information technologies;
- criminal actions of international groups, formations and individuals using quantum technologies and quantum information technologies.

The internal threats to the quantum resilience of national DPiBE are:

- illegal activities of political and economic structures in the field of using national DPiBE;
- violations of established regulations for the collection, processing and transmission of information in national DPiBE;
- intentional actions and unintentional errors of personnel of national DPiBE;
- disruption of technical means and software failures in national DPiBE.

The ways of influencing national DPiBE with the aim of violating quantum resilience are divided into informational, software and mathematical and physical.

The informational methods for violating the quantum resilience of national DPiBE include:

- intrusions of the targeting and timeliness of information exchange, illegal collection and use of information in national DPiBE using quantum technologies and/or quantum information technologies;
- unauthorized access to information resources of national DPiBE using quantum technologies and/or quantum information technologies;
- manipulation of information (disinformation, concealment or distortion of information) from national DPiBE using quantum technologies and/or quantum information technologies;
- illegal copying of data from national DPiBE using quantum technologies and/or quantum information technologies;
- destruction of information processing technology in national DPiBE using quantum technologies and/or quantum information technologies.

Software and mathematical methods for violating the quantum resilience of national DPiBE include:

- introduction of malware and viruses into national DPiBE using quantum technologies and/or quantum information technologies;
- installing software and hardware bugs into national DPiBE using quantum technologies and/or quantum information technologies;
- destruction or modification of data into national DPiBE using quantum technologies and/or quantum information technologies;
- defeat or destruction of information processing and communication facilities in national DPiBE using quantum technologies and/or quantum information technologies;
- destruction, disruption or theft of machine or other originals of information carriers;
- theft from national DPiBE using quantum technologies and/or quantum information technologies of software and/or hardware keys, means of cryptographic information protection;
- application of quantum information technologies and quantum algorithms for cryptographic analysis of information obtained from DPiBE;
- supply of components “infected” with the use of quantum technologies and/or quantum information technologies for use into national DPiBE.

As a result of the impact of quantum threats on national DPiBE, serious damage may be caused to the vital interests of the Russian Federation in political, economic, defense and other areas of state activity, and social and economic damage may be caused to society and individuals.

Methods and means of ensuring quantum resilience of national DPiBE

In order to prevent, counteract and neutralize quantum threats to national DPiBE, it is necessary to comprehensively apply legal, software and hardware, organizational and technical methods to ensure the quantum resilience of national DPiBE.

Legal methods for ensuring the quantum resilience of national DPiBE include the development of a set of regulatory legal acts, guidelines and regulatory and methodological documents on information protection in information systems and the use of quantum technologies and quantum information technologies. Considering the problems and risks associated with the use of quantum technologies and quantum information technologies, it is necessary to form in the field a new quantum legislation [31].

Software and hardware methods for ensuring the quantum resilience of national DPiBE include preventing leakage of processed information by eliminating unauthorized access to it, preventing special impacts that cause destruction, annihilation, distortion of information or failures in the operation of information technology, identifying embedded software or hardware errors, eliminating the interception of information by technical means, using cryptographic means of protecting information during transmission via communication channels, including using quantum key distribution and post-quantum cryptographic algorithms.

Organizational and economic methods for ensuring quantum resilience of national DPiBE include the formation and maintenance of systems for protecting confidential information in DPiBE using quantum information technologies, certification of these systems according to information security requirements, licensing of activities in the field of information security, standardization of methods and means of protecting information in DPiBE using quantum information technologies, control over the actions of personnel in protected DPiBE using intelligent methods (including quantum) and means.

An important place among these methods of ensuring the quantum resilience of the national DPiBE is occupied by motivation, economic incentives and psychological support for the activities of personnel involved in ensuring the quantum resilience of DPiBE, including the use of methods of multi-level filtering of potentially dangerous information and psychological influences [32].

Priority measures to ensure quantum resilience of national DPiBE

Priority measures to ensure quantum resilience of national DPiBE should include:

- development of forms, methods and means of implementing methods for ensuring quantum resilience of national DPiBE;
- preparation of decisions of executive authorities and documents that consolidate the main provisions of state policy to ensure the quantum resilience of national DPiBE;
- creation of a regulatory framework for implementing methods for ensuring quantum resilience of national DPiBE, including determining the sequence and procedure for developing legislative and regulatory legal acts, as well as mechanisms for the practical implementation of the adopted legislation;
- analysis of technical and economic parameters of domestic and foreign software and hardware for ensuring information security, including with the use of quantum information technologies and quantum key distribution, and the selection of promising areas for the development of domestic quantum technologies;
- formation of a scientific and technical program for the improvement and development of methods and means for ensuring quantum resilience of national DPiBE, providing for their use in national information and telecommunication networks and systems, taking into account the entry of the Russian Federation into global information networks and systems;
- creation of a certification system for domestic information technology tools for compliance with the requirements for ensuring quantum resilience;

- improving the organizational structure of the information security system of the Russian Federation in part of coordinating and regulating activities to ensure the security of the use of quantum information technologies and ensuring the quantum stability of DPiBE;
- development of a system of economic and statistical indicators characterizing the efficiency of DPiBE in the presence of quantum threats;
- determination of the real needs for specialists in ensuring the quantum resilience of national DPiBE, organization of a system for the selection, training and retraining of personnel.

Organizational framework for ensuring the resilience of national DPiBE

The results of the analysis of the state of quantum resilience of national DPiBE indicate the need to reform the existing organization of information security as a whole with the aim of integrating it into the information security system of the Russian Federation.

The organizational structure for ensuring quantum resilience of DPiBE of the Russian Federation consists of:

- state authorities and administration bodies of the Russian Federation and constituent entities of the Russian Federation solving problems within their competence;
- state and interdepartmental commissions and councils specializing in the development and implementation of digital platforms, blockchain ecosystems, quantum communications, quantum technologies, as well as ensuring quantum resilience and information security;
- research, design and engineering organizations involved in the development and implementation of digital platforms, blockchain ecosystems, quantum communications, quantum technologies, as well as ensuring quantum resilience and information security;
- operators of national DPiBE;
- educational institutions that train and retrain personnel for the development and implementation of digital platforms, blockchain ecosystems, quantum communications, quantum technologies, as well as ensuring quantum resilience and information security[16].

Main research areas within the project “Technologies for countering previously unknown quantum cyber threats”

The research group at the Sirius University of Science and Technology has already conducted and obtained the following results within the framework of the project “Technologies for countering previously unknown quantum cyber threats”:

1. Conceptual model of quantum security threats for DPiBE of the Russian Federation has been developed based on the addition and development of the well-known Methodology for assessing security threats by FSTEC of Russia and the Matrix of tactics and techniques of cyberattacks – MITRE Enterprise ATT&CK Matrix [33].
2. Based on the Kalman filter and the catastrophe theory of R.K. Thomas and V.I. Arnold, the mathematical model of the functioning of national DPiBE under the conditions of attacks by intruders using a quantum computer has been developed.

Currently, the common types of attacks on blockchain include, for example, [1, 34, 35]:

- 51% attack (a malicious actor controls more than 50% of the network’s processing power, allowing to manipulate the blockchain, potentially enabling double-spending or reversing transactions);
- Eclipse attack (a malicious actor attacks a single node of the blockchain network, creating an artificial false area around it, intercepting and replacing messages);
- Sybil attack (a malicious actor tries to capture and use a certain number of nodes of the blockchain network at once to generate incorrect data);
- Finney attack and Race attack (double-spending of funds in the blockchain system, if a miner accepts an unconfirmed transaction in the network);

[16] Currently, training and retraining in quantum stability of personnel related to ensuring quantum stability is carried out at the Sirius University of Science and Technology.

- Dust attack (formation of similar addresses in the network with the transfer of a small amount of money to the recipient's account in the hope that the next time the recipient will confuse the addresses and send the transfer to the wrong account);
- Denial of service (a malicious actor sends a large number of identical requests to a node of the blockchain network, DDoS);
- Cyberattacks on blockchain crypto primitives performed using quantum computers (quantum attacks).

3. Requirements have been developed for the system architecture, tools for ensuring quantum stability of national DPiBE using the cores of Enterprise Ethereum Alliance, Waves Enterprise, Hyperledger Fabric, Corda Enterprise, Masterchain, Microsoft Azure Blockchain in the context of a new quantum security threat.

The main areas of further research by the above-mentioned research group are:

- development of methods and models for ensuring verified security of national DPiBE in the context of a new quantum security threat;
- development of methods and algorithms for analyzing the resilience of national DPiBE in the context of a new quantum security threat based on modification of quantum algorithms (Shor, Grover, etc.);
- development of a methodology for solving problems of analyzing the quantum resilience of national DPiBE;
- development of methods and algorithms for parametric synthesis of quantum-resilience national DPiBE based on the theory of multi-criteria optimization in the context of a new quantum security threat;
- development of a methodology for solving problems of synthesizing technologies and programs for ensuring quantum resilience of national DPiBE;
- creation and development of the complex of logical and dynamic models to ensure the quantum resilience of national DPiBE;
- development of a software architecture for solving problems of analyzing (assessing) the resilience of national DPiBE under of a new quantum security threat;
- creation and debugging of a prototype of the software package for solving the problems of analysis (assessment) of the resilience of the operating of national DPiBE under of a new quantum security threat;
- detailing of contents and features of the implementation of the main stages of solving the problems of technology synthesis and comprehensive planning for ensuring quantum resilience of national DPiBE;
- defining the composition and structure of a possible analytical and simulation software package for synthesizing technology for ensuring quantum resilience of national DPiBE;
- detailing of composition and structure of mathematical and software support for solving the problems of analysis and synthesis of technologies and comprehensive plans for ensuring quantum resilience of national DPiBE;
- development of methods based on the integration of methods for ensuring verified security and the development of Agile and Waterfal approaches for designing quantum-resilience national DPiBE (Q-VSAWD);
- development based on the addition and development of methods of verified security and methodology of continuous development of digital platforms, taking into account security requirements, methodology for creating quantum-resilience national DPiBE (Q-DevSecOps);
- development of a software architecture for solving the problems of synthesizing quantum-resilience national DPiBE under of a new quantum security threat using methods of verified security and a methodology for continuous development taking into account security requirements;
- creation and debugging of a prototype of a software package for solving the problems of synthesizing quantum-resilience national DPiBE under of a new quantum security threat using methods of verified security and a methodology for continuous development taking into account security requirements.

This is not an exhaustive list of planned research aimed at implementing the proposed Concept.

Conclusion

The results obtained in the field of quantum informatics clearly demonstrate the high technological potential of quantum technologies. A cryptanalytically relevant or significant quantum computer can threaten the operating of various systems, including the functioning of national DPiBE in the Russian Federation.

The DPiBE of the Russian Federation do not have the required resilience of target operating under of attacks by intruders using quantum computers.

This article proposes and formulates the main provisions of the relevant Concept for ensuring the resilience of national DPiBE under a new quantum security threat. After discussion, coordination and approval, the Concept should become an integral part of the Information Security Concept of the Russian Federation.

In developing the Concept, were summarized, systematized and comprehensively rethought the results obtained in the works well-known works of Russian scientists on quantum information technologies and the results obtained in the works of the authors of this article and other members of the group "Technologies for countering previously unknown quantum cyber threats" of the Scientific Center for Information Technology and Artificial Intelligence of the Sirius University of Science and Technology.

This article also reviewed the main directions of further research of the above-mentioned group, aimed at implementing the proposed Concept.

REFERENCES

1. **Ishchukova E.A., Panasenko S.P., Romanenko K.S., Salmanov V.D.** *Kriptograficheskie osnovy blokchein-tekhnologii* [*Cryptographic foundations of blockchain technologies*]. Moscow: Izdatel'stvo «DMK Press», 2022. 301 p.

2. **Petrenko A.S.** *Kvantovo-ustoichivyi blokchein. Kak obespechit' bezopasnost' blokchein-ekosistem i platform v usloviiakh atak s ispol'zovaniem kvantovogo komp'iutera* [*Quantum-resilience blockchain: How to ensure the security of blockchain ecosystems and platforms in the face of attacks using a quantum computer*]. St. Petersburg: Piter, 2023. 320 p.

3. **Zaborovsky V.S., Lei Zhang, Skiba V.Yu., Strekalov S.V.** Digital information and logistic platform for operational management of foreign trade activities of high-tech product suppliers. *St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunications and Control Systems*, 2018, Vol. 11, No. 4, Pp. 7–20. DOI: 10.18721/JCSTCS.11401

4. **Skiba V.Yu., Petrenko S.A., Murzina A.A., Popova K.R.** New types of threats and assessment of quantum stability of information systems in the field of foreign trade activity. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 4, Pp. 16–34. DOI: 10.18721/JCSTCS.17402

5. **Petrenko S.A., Petrenko A.S., Ozhiganova M.I.** Concept of ensuring the cyber resilience of the Bank of Russia digital ruble platform in the face of growing security threats. *Zaŝita informacii. Inside*, 2024, Vol. 119, No. 5, Pp. 6–13.

6. **Terekhov A.I.** On the Development of the Scientific Base of Quantum Technologies. *Economics of Science*, 2022, Vol. 8, No. 1, Pp. 58–72. DOI: 10.22394/2410-132X-2022-8-1-58-72

7. **Terekhov A.I.** Bibliometric Analysis of Academic Literature on Quantum Information Processing. *Photonics*, 2024, Vol. 18, No. 4, Pp. 296–312. DOI: 10.22184/1993-7296.FRos.2024.18.4.296.312

8. **Skiba V.Yu., Petrenko S.A., Murzina A.A., Popova K.R.** Evaluation of quantum stability of information systems of customs authorities of the Russian Federation in present-day conditions. *Vestnic of Russian Customs Academy*, 2024, Vol. 69, No. 4, Pp. 32–45.

9. **Stupin D.D., Petrenko A.S., Petrenko S.A.** Razvitie tekhnologii kvantovykh vychislenii i sviazannye s nim ugrozy dlia kriticheskoi informatsionnoi infrastruktury Rossiiskoi Federatsii [Development of quantum

computing technologies and associated threats to the critical information infrastructure of the Russian Federation]. *XVI Vserossiiskaia Mul'tikonferentsiia po Problemam Upravleniia* (*MKPU-2023*) [*XVI All-Russian Multi-Conference on Management Problems* (*MCMP-2023*)], 2023, Pp. 168–172.

10. **Kazarin O.V., Skiba V.Yu., Sharyapov R.A.** Novye raznovidnosti ugroz mezhdunarodnoi informatsionnoi bezopasnosti [New types of threats to international information security]. *RSUH/RGGU BULLETIN "Information Science. Information Security. Mathematics" Series*, 2016, Vol. 3, No. 1, Pp. 54–72.

11. **Konyavsky V.A., Ross G.V., Sychev A.M., Skiba V.U.** Information protection in systems of critical information infrastructures. *Journal of the Balkan Tribological Association*, 2021, Vol. 27, No. 4, Pp. 479–496.

12. **Skiba V.Yu., Turgiev E.Z., Lisov D.N., Salokina (Polyakova) N.A., Sergeev V.S.** Proaktivnaia bezopasnost' kvantovykh AIS [Proactive Security of Quantum Automated Information Systems]. *VIII International Conference "The 2024 Symposium on Cybersecurity of the Digital Economy – CDE'24"*, 2025, Pp. 71–77.

13. **Petrenko S.A., Petrenko A.S.** O protivodeistvii ranee neizvestnym kvantovym kiberugrozam [On countering previously unknown quantum cyber threats]. *VII International Conference "The 2023 Symposium on Cybersecurity of the Digital Economy – CDE'23"*, 2024, Pp. 13–23.

14. **Shor P.W.** Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 1997, Vol. 26, No. 5, Pp. 1484–1509. DOI: 10.1137/S0097539795293172

15. **Shor P.W.** Algorithms for quantum computation: discrete logarithms and factoring. Proceedings *35th Annual Symposium on Foundations of Computer Science*, 1994, Pp. 124–134. DOI: 10.1109/SFCS.1994.365700

16. **Simon D.R.** On the power of quantum computation. *Proceedings 35th Annual Symposium on Foundations of Computer Science*, 1994, Pp. 116–123. DOI: 10.1109/SFCS.1994.365701

17. **Holevo A.S.** *Vvedenie v kvantovuiu teoriiu informatsii* [*Introduction to Quantum Information Theory*]. Moscow: MTsNMO, 2002. 128 p.

18. **Holevo A.S.** *Veroiatnostnye i statisticheskie aspekty kvantovoi teorii* [*Probabilistic and statistical aspects of quantum theory*]. Moscow: MTsNMO; NMU, 2020. 364 p.

19. **Bogdanov Y.I., Valiev K.A., Kokin A.A.** Quantum computers: achievements, implementation difficulties, and prospects. *Microelectronics*, 2011, Vol. 40, No. 4, Pp. 243–255. DOI: 10.1134/S1063739711040032

20. **Nielsen M.A., Chang I.L.** *Quantum Computation and Quantum Information: 10th Anniversary ed*. Cambridge: Cambridge University Press, 2011. 702 p.

21. **Bennett C.H., Shor P.W.** Quantum information theory. *IEEE Transactions on Information Theory*, 1998, Vol. 44, No. 6, Pp. 2724–2742. DOI: 10.1109/18.720553

22. **Arute F., Arya K., Babbush R. et al.** Quantum supremacy using a programmable superconducting processor. *Nature*, 2019, Vol. 574, Pp. 505–510. DOI: 10.1038/s41586-019-1666-5

23. **Petrenko S.A., Petrenko A.S., Kostyukov A.D.** Countermeasures technologies previously unknown quantum cyber threats. *Zaŝita informacii. Inside*, 2024, Vol. 118, No. 4, Pp. 66–76.

24. **Petrenko A.S., Petrenko S.A.** Quantum resilience estimation method blockchain. *Cybersecurity Issues*, 2022, Vol. 49, No. 3, Pp. 2–22. DOI: 10.21681/2311-3456-2022-3-2-22

25. **Mosca M.** Cybersecurity in an era with quantum computers: Will we be ready? *IEEE Security & Privacy*, 2018, Vol. 16, No. 5, Pp. 38–41. DOI: 10.1109/MSP.2018.3761723

26. **Skiba V.Yu.** *Ob"ektno-funktsional'naia verifikatsiia informatsionnoi bezopasnosti raspredelennykh avtomatizirovannykh informatsionnykh sistem tamozhennykh organov. Diss. doktora tekhn. nauk* [*Object-functional verification of information security of distributed automated information systems of customs authorities. Doctor of Technical Sciences diss.*]. St. Petersburg, 2009. 365 p.

27. **Moldovyan N.A., Petrenko A.S.** Algebraic signature algorithms with two hidden groups. *Cybersecurity Issues*, 2024, Vol. 64, No. 6, Pp. 98–107. DOI: 10.21681/2311-3456-2024-6-98-107

28. **Vysotskaya V.V., Chizhov I.V.** The security of the code-based signature scheme based on the Stern identification protocol. *Applied Discrete Mathematics*, 2022, Vol. 57, Pp. 67–90. DOI: 10.17223/20710410/57/5

29. **Vasyutina A.P., Klyucharev P.G.** Optimization of a post-quantum cryptographic protocol based on isogeny of supersingular elliptic curves. *Bezopasnye Informatsionnye Tekhnologii* [*Secure Information Technologies*], 2023, Pp. 40–43.

30. **Wang R., Dubrova E.** A shared key recovery attack on a masked implementation of CRYSTALS-kyber's encapsulation algorithm. In: *Foundations and Practice of Security. FPS 2023* (eds. M. Mosbah, F. Sèdes, N. Tawbi, T. Ahmed, N. Boulahia-Cuppens, J. Garcia-Alfaro), 2024, Vol. 14551, Pp. 424–439. DOI: 10.1007/978-3-031-57537-2_26

31. **Gromova E.A., Petrenko S.A.** Quantum law: The beginning. *Journal of Digital Technologies and Law*, 2023, Vol. 1, No. 1, Pp. 62–88. DOI: 10.21202/jdtl.2023.3

32. **Gnidko K.O., Sadreev K.R., Lisov D.N.** Kontseptsiia povysheniia ustoichivosti avtomatizirovannoi sistemy k novym ugrozam tipa otkaz v obsluzhivanii cheloveka [The concept of increasing the resilience of an automated system to new threats such as denial of human service]. *VII International Conference "The 2023 Symposium on Cybersecurity of the Digital Economy – CDE'23"*, 2024, Pp. 71–74.

33. **Petrenko S.A., Balyabin A.A.** A model of quantum threats to information security for national blockchain ecosystems and platforms. *Cybersecurity Issues*, 2025, Vol. 65, No. 1, Pp. 7–17. DOI: 10.21681/2311-3456-2025-1-7-17

34. **Saha B., Hasan M.M., Anjum N., Tahora S., Siddika A., Shahriar H.** Protecting the decentralized future: An exploration of common blockchain attacks and their countermeasures. *arXiv:2306.11884*, 2023. DOI: 10.48550/arXiv.2306.11884

35. **Guru A., Mohanta B.K., Mohapatra H., Al-Turjman F., Altrjman C., Yadav A.** A survey on consensus protocols and attacks on blockchain technology. *Applied Sciences*, 2023, Vol. 13, No 4, Art. no. 2604. DOI: 10.3390/app13042604

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Skiba Vladimir Yu.
Скиба Владимир Юрьевич
E-mail: vskiba69@mail.ru
ORCID: https://orcid.org/0000-0002-9805-7800

Petrenko Sergei A.
Петренко Сергей Анатольевич
E-mail: petrenko.sa@talantiuspeh.ru
ORCID: https://orcid.org/0000-0003-0644-1731

Gnidko Konstantin O.
Гнидко Константин Олегович
E-mail: gnidko.ko@talantiuspeh.ru
ORCID: https://orcid.org/0000-0002-8605-8865

Petrenko Alexei S.
Петренко Алексей Сергеевич
E-mail: a.petrenko1999@rambler.ru
ORCID: https://orcid.org/0000-0002-9954-4643

Submitted: 20.03.2025; Approved: 12.05.2025; Accepted: 30.05.2025.
Поступила: 20.03.2025; Одобрена: 12.05.2025; Принята: 30.05.2025.

Research article
DOI: https://doi.org/10.18721/JCSTCS.18206
UDC 004.89

IT PROJECT INFRASTRUCTURE SETUP AUTOMATION WITH HELP OF LARGE LANGUAGE MODELS

V.A. Ivlev ✉ , *I.V. Nikiforov* , *S.M. Ustinov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ nevidd@yandex.ru

Abstract. This study conducts an analysis of existing large language models (LLMs) and AI agents, identifying Llama 2 as the most suitable model for automating IT project environment configuration. A mathematical model of the proposed method is introduced to automate IT infrastructure setup and reduce development time. The system architecture incorporates modules for natural language processing (NLP), configuration generation and command execution. The effectiveness of the method is evaluated through experiments across five key production scenarios, comparing two approaches: traditional infrastructure configuration tools and the proposed LLM-based method utilizing Llama 2. Experimental results demonstrate that the proposed method reduces configuration time up to 60%, decreases error rates from 25% to 8% and improves configuration quality approximately in 3 times. The article is relevant to IT professionals engaged in automating development and infrastructure configuration processes, as well as researchers exploring the application of artificial intelligence, particularly large language models, in the IT industry.

Keywords: Large Language Model, Llama 2, AI agent, IT infrastructure setup automation, Natural Language Processing, configuration generation, artificial intelligence

Citation: Ivlev V.A., Nikiforov I.V., Ustinov S.M. IT project infrastructure setup automation with help of large language models. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 74–90. DOI: 10.18721/JCSTCS.18206

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18206
УДК 004.89

АВТОМАТИЗАЦИЯ НАСТРОЙКИ ИНФРАСТРУКТУРЫ ИТ-ПРОЕКТА С ИСПОЛЬЗОВАНИЕМ LLM-МОДЕЛЕЙ

В.А. Ивлев ✉ , *И.В. Никифоров* , *С.М. Устинов*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ nevidd@yandex.ru

Аннотация. В исследовании проведен анализ существующих больших языковых моделей (LLM) и AI-агентов, на основе которого выбрана модель Llama 2 как наиболее подходящая для автоматизации настройки окружения ИТ-проекта. Предложена математическая модель метода, позволяющего автоматизировать процесс настройки ИТ-инфраструктуры и сократить время разработки ИТ-проекта. Разработана архитектура системы, включающая модули для обработки естественного языка (NLP), генерации конфигураций и выполнения команд. Оценена эффективность предложенного метода в экспериментах на пяти основных производственных сценариях. В ходе экспериментов сравнивались два подхода настройки ИТ-инфраструктуры: подход с использованием традиционных средств настройки инфраструктуры и подход с использованием предложенного в работе метода на основе LLM-модели Llama 2. Показано, что использование предложенного метода позволяет сократить время настройки до 60%, снизить количество ошибок с 25% до 8% и повысить качество настройки приблизительно в 3 раза. Статья представляет интерес для специалистов в области информационных технологий, занимающихся автоматизацией процессов разработки и настройки инфраструктуры, а также для исследователей, изучающих применение искусственного интеллекта, а именно больших языковых моделей, в ИТ-индустрии.

Ключевые слова: большая языковая модель, Llama 2, AI-агент, автоматизация настройки ИТ-инфраструктуры, обработка естественного языка, генерация конфигураций, искусственный интеллект

Для цитирования: Ivlev V.A., Nikiforov I.V., Ustinov S.M. IT project infrastructure setup automation with help of large language models // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 74–90. DOI: 10.18721/JCSTCS.18206

Introduction

Nowadays software development plays a key role in many companies, as it enables the automation of processes, improves the quality of production lines and enhances the efficiency of management processes [1, 2]. At the same time, the software being developed grows larger in scale and more complex with each new task. As software complexity increases, companies face new challenges and management tasks related to configuring IT systems, aimed at maintaining software development and operational processes. This results in developers encountering a growing demand for effective tools to automate the configuration of IT infrastructure.

To address this problem, companies are trying to adopt methods, algorithms, approaches and systems capable of automating the configuration of IT project infrastructure. An optimal solution in this context could be a system that interprets human-readable task descriptions into command formats and automatically configures IT infrastructure by executing these commands. However, identifying or developing such a system is a non-trivial task. Many commercially available systems are proprietary and require skilled personnel [1, 3] to maintain or implement them.

To solve the problem of automating IT project infrastructure configuration a method based on a Natural Language Processing (NLP) model service was previously proposed. This method automates infrastructure configuration thereby accelerating various stages of IT project development [1]. The key component of this approach is its ability to interpret human-readable descriptions (unformalized text) into executable command sequences, which are then executed by the system to configure the IT infrastructure. However, the application of Large Language Models (LLMs) for NLP automation was not addressed, despite their potential to enhance the degree of automation and output quality.

To overcome this limitation, we propose an approach that employs an LLM as an interpreter to convert human-readable descriptions into executable commands. This serves as an implementation of the NLP architectural block within the method for automating IT infrastructure configuration in IT projects [3].

Relevance of the topic

Modern IT projects typically involve complex and diverse technologies [2], services and platforms [4, 5], including software and hardware components, such as automated workstation equipment, virtualization and containerization systems, infrastructure monitoring platforms, system-wide software (operating systems, Database Management System (DBMS), core infrastructure service software etc.), telephony and videoconferencing systems, data center engineering infrastructure, cybersecurity subsystems and others. Configuring such infrastructure is a labor-intensive task requiring specialized knowledge and expertise. Automating the deployment of IT infrastructure could potentially simplify environment configuration, thereby enhancing development productivity and reducing the time required to deliver the final product. Integrating tools for code generation based on natural language descriptions could further streamline this process, virtually eliminating the need for direct programmer involvement. Automation of IT infrastructure configuration using tools like LLMs has the potential to significantly reduce time [6]. LLMs can rapidly analyze project requirements formulated in natural language, propose optimal configurations and generate code for environment setup. A configuration process leveraging LLMs or LLM-based tools can minimize human errors associated with infrastructure setup, leading to more stable and reliable project operations.

LLMs are trained on huge datasets [7] enabling them to adapt to diverse use cases [8] and improve their performance over time [9]. This adaptability is particularly valuable in dynamic IT environments. Such LLM-driven systems can be applied to various aspects of IT infrastructure configuration, including parameter optimization [10], network setup, security management etc. This makes the natural-language-based code generation approach flexible and scalable for different project types. LLMs effectively generate code because of their big volume training data [11, 12], which incorporates code examples and technical documentation, allowing them to comprehend the syntax and logic of multiple programming languages [13]. They analyze query context to ensure solution accuracy and relevance and can interact with users to clarify requirements. Beyond code writing, LLMs can generate tests and documentation [14], positioning them as indispensable tools for developers [15]. Solutions employing LLMs to execute business-oriented tasks are termed AI agents.

Thus, the use of AI agents [16] for automating IT infrastructure configuration represents a promising approach that could substantially enhance the efficiency and reliability of IT projects. By automating operational tasks, this strategy frees human resources to focus on tactical and strategic challenges.

Degree of development of the topic and analysis of existing AI agents

Given the relevance of utilizing tools for code generation [17] based on natural language descriptions as a core module [18] of the proposed automation system, a study of existing AI agents and their underlying LLMs is conducted. Considering the specificity of the domain, a list of candidates (Table 1) potentially capable of solving code generation tasks from natural language descriptions is identified. It

Table 1

Comparative analysis of existing natural language-driven code generation tools

AI agent (LLM-model)	Model type	Model size	Architecture	Security and robustness	Broad applicability	Software integration	Open source	Multimodality	Computational optimization
Claude-3Opus	Model from Anthropic	Not disclosed	Transformer [19]	+	+	-	-	+	-
Gemini-1Ultra [20]	Model from Google	Not disclosed	Transformer	+	+	-	-	+	+
Mistral-Large	Model from Mistral	12.9B	Sparse Mixture of Experts	-	+	-	+	-	+
Llama 2 [21]	Model from Microsoft	7B, 13B, 70B	Transformer	-	+	+	+	-	+
Qwen-1.5	Model from Huawei	7B, 13B	Transformer	-	+	-	-	+	+
DeepSeek	Model from DeepSeek	Not disclosed	Transformer	-	+	+	-	-	+
Baichuan-2 Turbo	Model from Baichuan	7B, 13B	Transformer	-	+	-	-	-	+
Copilot	Model from OpenAI	Not disclosed	Transformer	-	+	+	-	-	+
Codex	Model from OpenAI	Not disclosed	Transformer	-	+	+	-	-	+
FauxPilot	Model from Replit	Not disclosed	Transformer	-	+	+	-	-	+
StarCoder	Model from BigCode	Not disclosed	Transformer	-	+	+	-	-	+
GPT-4-Turbo [22]	Model from OpenAI	Not disclosed	Transformer	+	+	+	-	-	+
EleutherAI GPT	Model from EleutherAI	1.3B, 2.7B, 6B, 20B	Transformer	-	+	-	+	-	+
Microsoft Turing	Model from Microsoft	Not disclosed	Transformer	+	+	-	-	-	+
IBM Watson NLU	Model from IBM	Not disclosed	Diverse NLP Models	+	+	-	-	-	+

is worth noting that alongside classical approaches, the Retrieval-Augmented Generation (RAG) methodology combines generative models with data retrieval from external sources to enhance code accuracy. However, its integration with LLMs necessitates a dedicated analysis of architectural considerations. This study focuses on evaluating the "pure" generative capabilities of the models, deferring the investigation of hybrid RAG-based systems to future work.

The study proposed limiting the comparative analysis to the most critical criteria for addressing IT environment configuration automation, namely: security and robustness of results, breadth of model applicability across diverse IT project domains, integration capabilities with third-party software, open-source availability, multimodality and optimization for internal computations within the model.

It should be noted that all analyzed AI agents and LLMs adequately account for and use context during task execution for code generation or NLP, where understanding broad context [23] can significantly enhance result quality [24]. At first glance, LLMs such as GPT-4-Turbo, Gemini-1Ultra and Microsoft Turing demonstrate broad applicability and can be employed for diverse tasks, ranging from text generation to integration with cloud services and business solutions. These models are universal and adaptable to a wide range of challenges. AI agents like Codex, Copilot and StarCoder are specialized in software development support, making them indispensable tools for programmers engaged in code automation and software solution creation. However, these criteria for LLMs and AI agents are of lesser significance as the listed tools lack open-source availability, precluding their modification and adaptation to task-specific requirements.

Thus, based on the conclusions drawn from the initial evaluation of LLMs and AI agents, the focus should shift to open-source tools such as Llama 2 [21], EleutherAI GPT and Mistral-Large. These provide researchers and developers with the ability to modify and adapt models for specialized tasks [25], which is critical for projects requiring model fine-tuning [26] for use in niche domains [27]. Notably, unlike EleutherAI GPT and Mistral-Large, Llama 2 offers software integration – specifically, functionality focused on programming assistance and code generation [28]. Considering this criterion alongside the aforementioned conclusions, Llama 2 considered as the preferred candidate for implementation as the script generation module in a system designed to automate IT infrastructure configuration for IT projects.

A method for automating IT project infrastructure configuration through the application of an LLM

The proposed method is formalized through the following parameters and objective function.

1. Method parameters:

D: set of all human-readable task descriptions;

C: set of all technical configurations;

$F(d)$: transformation function from task description $d \in D$ to configuration $c \in C$;

ε: probability that $F(d)$ is an incorrect configuration.

The set of valid technical configurations C_{corr} is defined as a subset C, where the probability of successful transformation exceeds a predefined threshold $1 - \varepsilon$:

$$T_{corr} = \left\{ c \in C \middle| \exists d \in D : F\left(d\right) = c \wedge P_{corr}\left(c \middle| d\right) \geq 1 - \varepsilon \right\},$$

where $P_{corr}(c|d)$ denotes the conditional probability, that configuration c generated from task description d, is correct.

2. Objective function – the target function minimizes infrastructure configuration time T_{auto}:

$$T_{auto} = T_{Llanna} + T_{exec},$$

where T_{Llanna} is the time spent generating the configuration; T_{exec} is the time required to execute the script on the actual infrastructure.

3. Model quality – model performance is evaluated using the accuracy metric:

$$Accuracy = \frac{\left|F\left(d\right) \cap F_{valid}\right|}{\left|D\right|},$$

where F_{valid} is the set of correct configurations.

Objective of the mathematical model – the proposed mathematical model aims to optimize the IT infrastructure automation process using Llama 2 LLM. Specifically, it focuses on:

$$M = \min\left(T_{auto}\right) \wedge \max\left(Accuracy\right) \wedge \min\left(\varepsilon\right),$$

that is, minimizing execution time (T_{auto}), increasing the accuracy of converting human-readable descriptions into valid configurations ($Accuracy$), reducing error probability (ε) and ensuring system stability. By effectively combining configuration generation and execution the model reduces reliance on software developer intervention, thereby enhancing productivity [30] and infrastructure configuration reliability. This method forms the foundation of the software system developed in this study to address automation challenges.

Implementation of the IT infrastructure automation method

To integrate the Llama 2 LLM into the IT infrastructure automation process, a dedicated architecture is designed (Fig. 1). This architecture comprises several core components aimed at efficiently converting human-readable task descriptions into executable infrastructure configuration commands.

The architectural elements are analyzed in detail below.

1. *Data input module*. This module is responsible for receiving and preprocessing human-readable task descriptions. Descriptions may be provided as text files, API requests or via user interface. The module also performs text normalization, stop-word removal and tokenization to prepare data for model processing.

As the technology stack of this module for user interaction, the following is used: user interaction is implemented via a REST interface using the Spring Boot 3 framework, enabling rapid deployment of a reliable server-side application. Input data validation (task descriptions) adheres to the Java API Bean Validation specification (JSR 380). Task storage is managed in PostgreSQL database using Spring Data JPA. For CLI integration, the Picocli tool is employed alongside the Spring Shell framework.

2. *NLP module*. This module leverages the Llama 2 LLM [31] to analyze and interpret textual descriptions. The model converts text requests into structured data for configuration generation and may request user clarification for ambiguous inputs.

The following is used as the technology stack: integration with Llama 2 is achieved via REST interface or gRPC using Spring WebClient or gRPC-Java libraries. Text preprocessing (tokenization, entity extraction) utilizes Apache OpenNLP, augmented with custom rules. Complex scenarios employ spaCy's language parser via Java bindings.

3. *Configuration generation module*. Using data from the NLP module, this module generates executable commands or scripts for infrastructure configuration. Outputs include configuration files [32], service deployment commands and network parameter setups. Generated configurations undergo validation before execution.

The following is used as the technology stack: configuration files (YAML, JSON, scripts) are templated using tools Apache Velocity or Thymeleaf.

4. *Command execution module*. This module executes generated commands on target infrastructure, such as cloud platforms, servers or containerized environments. It monitors execution status and provides feedback.

The following is used as the technology stack: secure SSH connections are established via library JSch. Asynchronous task handling uses CompletableFuture and Project Reactor. Cloud orchestration leverages Kubernetes Java client. Parallel task management employs ThreadPoolExecutor class or reactive streams.

5. *Monitoring and feedback module*. Post-execution, this module collects configuration results, including errors and warnings [33] to refine future configurations and improve model performance.

The following is used as the technology stack: metrics are exported to system Prometheus and Grafana via Micrometer tool. Logging is implemented with Log4j2 library, integrated into an ELK stack (Elasticsearch, Logstash, Kibana) [34]. Notifications are dispatched via Spring Integration framework, supporting tools like Slack, E-mail and Telegram. It is important to note that, the developed tool does not always produce fully accurate automation scripts. Generated outputs serve as templates that users may manually refine to meet specific requirements.

Example of method application

To provide a clear visualization of the method's workflow a diagram has been developed (Fig. 2), illustrating the interaction between the system's core modules. The diagram encompasses the following stages.

1. Data input. The user submits a textual task description via an interface or API.

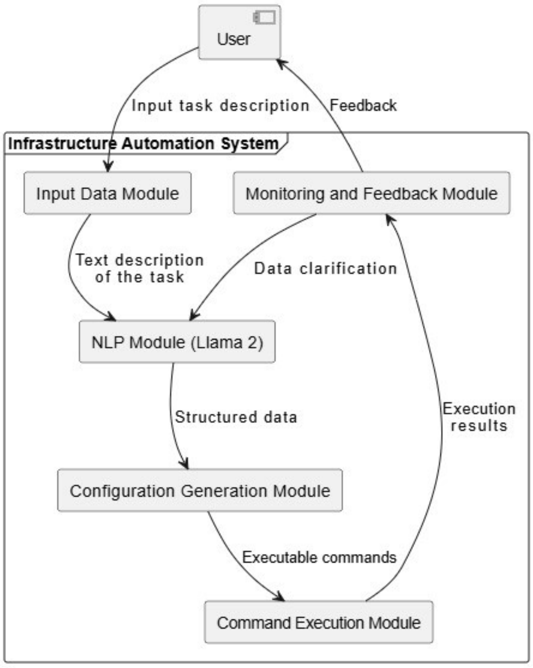


Fig. 1. Component diagram of the automation system with Llama 2 LLM integration

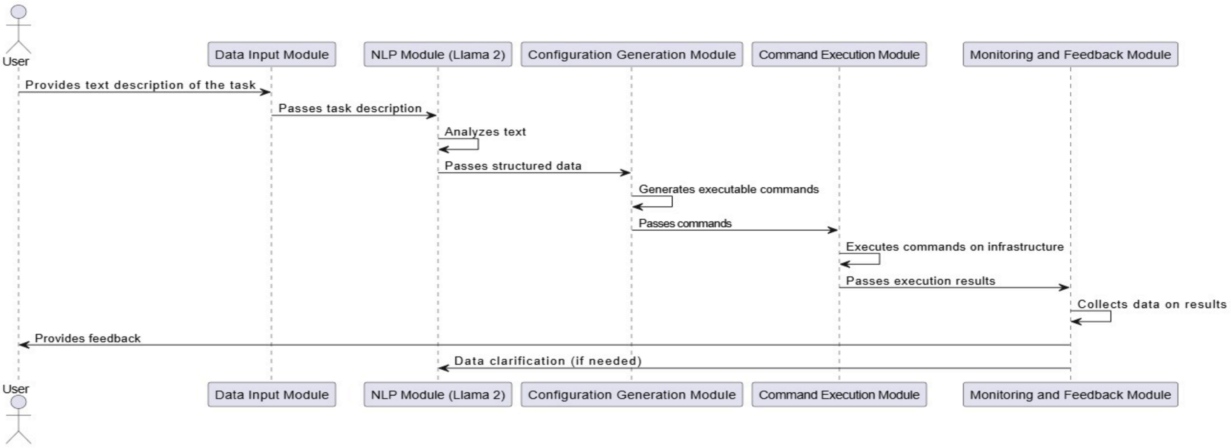


Fig. 2. MSC diagram of system component message exchange in the demonstration scenario

2. NLP. The Llama 2 model parses the text and converts it into structured data.

3. Configuration generation. Executable commands are generated based on the data derived from the model [35, 36].

4. Command execution. The commands are deployed on the target infrastructure.

5. Monitoring and feedback. The system collects execution results and delivers feedback to the user.

To demonstrate the method's application, consider an example of configuring infrastructure for a PostgreSQL database replication project (Fig. 3). Below is a flowchart of the method's workflow with each stage of the automation system's operation explained in detail.

1. *Data input*. The user submits the task description: "Configure PostgreSQL with replication across two servers: a primary server and a backup server." An example of system logging at this stage is provided below (listing 1).

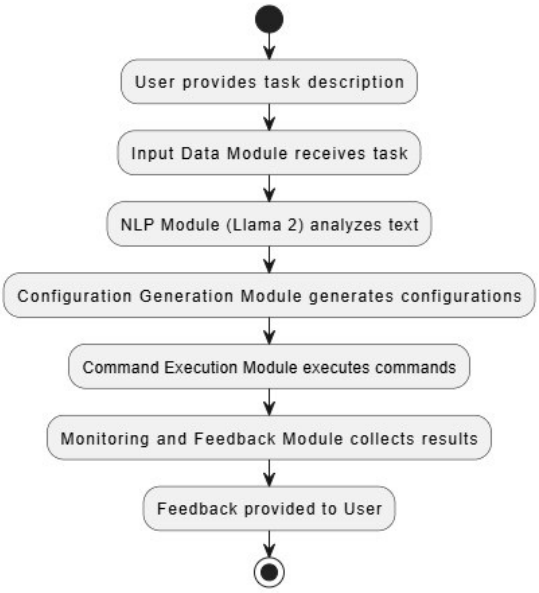

User provides task description
Input Data Module receives task
NLP Module (Llama 2) analyzes text
Configuration Generation Module generates configurations
Command Execution Module executes commands
Monitoring and Feedback Module collects results
Feedback provided to User

Fig. 3. Flowchart of the method's workflow

Listing 1. *Example of system logging during the data input stage*.

```
[INFO] [2025-02-20 21:59:45] User provided the task description:
"Set up PostgreSQL with replication on two servers: primary server and backup server."
```

2. *NLP*. The Llama 2 model analyzes the text and identifies key parameters: database type (PostgreSQL), server count (2) and replication mode. An example of system logging at this stage is provided below (listing 2).

Listing 2. *Example of system logging during the NLP processing stage*.

```
[INFO] [2025-02-20 22:00:03] Llama 2 model started text analysis.
[DEBUG] [2025-02-20 22:00:31] Identified key parameters:
- Database type: PostgreSQL
- Number of servers: 2
- Replication mode: master-slave
[INFO] [2025-02-20 22:00:35] Text analysis completed successfully.
```

3. *Configuration generation*. Based on the model's analysis, commands are generated to configure the PostgreSQL database (listing 3), including primary server setup, replication parameters and backup server creation.

Listing 3. *Example of a Llama 2 model-generated command for PostgreSQL database configuration*.

```
# Primary Server Configuration
sudo sed -i "s/#wal_level = replica/wal_level = replica/" /etc/postgresql/12/main/postgresql.conf
sudo sed -i "s/#max_wal_senders = 3/max_wal_senders = 3/" /etc/postgresql/12/main/postgresql.conf
sudo sed -i "s/#hot_standby = on/hot_standby = on/" /etc/postgresql/12/main/postgresql.conf
sudo sed -i "s/#hot_standby = on/hot_standby = on/" /etc/postgresql/12/main/postgresql.conf
sudo -u postgres psql -c "CREATE ROLE replicator WITH REPLICATION LOGIN PASSWORD
'*******';"
sudo systemctl restart postgresql

# Backup Server Configuration
sudo pg_basebackup -h primary_server -D /var/lib/postgresql/12/main -U replicator -P -v

echo "standby_mode = 'on'
primary_conninfo = 'host=primary_server port=5432 user=replicator password=******* '" | sudo tee /
var/lib/postgresql/12/main/recovery.conf

sudo chown -R postgres:postgres /var/lib/postgresql/12/main
sudo systemctl restart postgresql
```

An example of the automation system’s console output at this stage is provided below (listing 4).

Listing 4. *Example of system logging during the configuration generation stage*.

```
[INFO] [2025-02-20 22:01:24] Configuration generation started.
[DEBUG] [2025-02-20 22:01:37] Generated commands for primary server (master) setup:
1. Configuration of `postgresql.conf`:
  - wal_level = replica
  - max_wal_senders = 3
  - hot_standby = on
2. Creation of replication user:
  - CREATE ROLE replicator WITH REPLICATION LOGIN PASSWORD *******;
[DEBUG] [2025-02-20 22:02:21] Generated commands for backup server (replica) setup:
1. Configuration of `recovery.conf`:
  - standby_mode = on
  - primary_conninfo = 'host=primary_server port=5432 user=replicator password=*******'
2. Initialization of the backup server:
  - pg_basebackup -h primary_server -D /var/lib/postgresql/12/main -U replicator -P -v
[INFO] [2025-02-20 22:03:07] Configuration generation completed successfully.
```

4. *Command execution*. The commands are executed on the target servers. The primary server is configured as a master and the backup server as a replica. An example of the automation system’s console output at this stage is provided below (listing 5).

Listing 5. *Example of system logging during the command execution stage*.

```
[INFO] [2025-02-20 22:07:34] Command execution started on the target server (primary server).
[DEBUG] [2025-02-20 22:07:53] Commands executed on the primary server:
1. `postgresql.conf` parameters updated successfully.
2. User `replicator` created.
[INFO] [2025-02-20 22:08:06] Command execution started on the target server (backup server).
[DEBUG] [2025-02-20 22:08:21] Commands executed on the backup server:
1. `recovery.conf` parameters updated successfully.
2. Backup server initialized using `pg_basebackup`.
[INFO] [2025-02-20 22:08:44] Command execution completed successfully.
```

5. *Monitoring and feedback*. The system verifies the replication status and delivers a report to the user: “Replication configured successfully. Primary server: active. Backup server: synchronized.” An example of the automation system’s console output at this stage is provided below (listing 6).

Listing 6. *Example of system logging during the monitoring and feedback stage*.

```
[INFO] [2025-02-20 22:09:02] Replication status monitoring started.
[DEBUG] [2025-02-20 22:09:11] Replication status check:
- Primary server: active, replication enabled.
- Backup server: synchronized with the primary server.
[INFO] [2025-02-20 22:09:49] Monitoring completed successfully.
[INFO] [2025-02-20 22:09:51] Report provided to the user:
"Replication successfully configured. Primary server: active. Backup server: synchronized."
[INFO] [2025-02-20 22:09:52] Task "Set up PostgreSQL with replication" completed successfully.
```

Experimental results

The experiment aimed to compare the proposed approach with traditional configuration tools. Traditional automation tools include: integrated development environments (IDEs), web configurators, administrative panels for information systems, bash/batch scripts, CLI utilities, manual scripting, GUI tools and others [37].

Key task-specific metrics are selected for comparison, including development time, error correction time, error counts at various stages (compilation, code review, others) and overall solution quality and efficiency.

The experiment involved five distinct scenarios, each representing a standard infrastructure configuration task. These scenarios are chosen to cover core infrastructure components applicable to most IT projects:

1. PostgreSQL database configuration with primary and backup server replication.
2. Kubernetes setup with automatic scaling.
3. CI/CD pipeline configuration using GitLab and Jenkins.
4. Load balancer configuration across servers.
5. Monitoring system setup with Prometheus and Grafana metrics visualization [34].

System pre-configuration was performed on a hardware platform comprising an Intel Xeon E5-2666 v3 processor (2.90 GHz), 32 GB of RAM and an NVIDIA GeForce RTX 3050 Ti graphics processing unit (GPU) with 8 GB of dedicated memory. Preparing the baseline environment – including the installation and configuration of Java and Docker, subsequent deployment of the software implementing the proposed approach, installation of dependencies, runtime environment setup for the Llama 2 model, its initialization, and functional testing – required approximately 6–8 hours of a software engineer’s effort. Such configuration steps, essential for ensuring software compatibility and operational stability, are excluded from the experimental results, as they represent a one-time infrastructure preparation activity.

Each scenario is executed by three developers of varying expertise: junior, middle and senior engineers. This design allowed assessing how the method’s efficiency varies with developer experience. The experimental data included:

- average lines of code (LOC) per scenario;
- time spent on configuration using traditional tools;
- time spent on automated development and configuration;
- error counts during compilation, code review and other stages;
- solution quality and efficiency, measured as the ratio of LOC to error count.

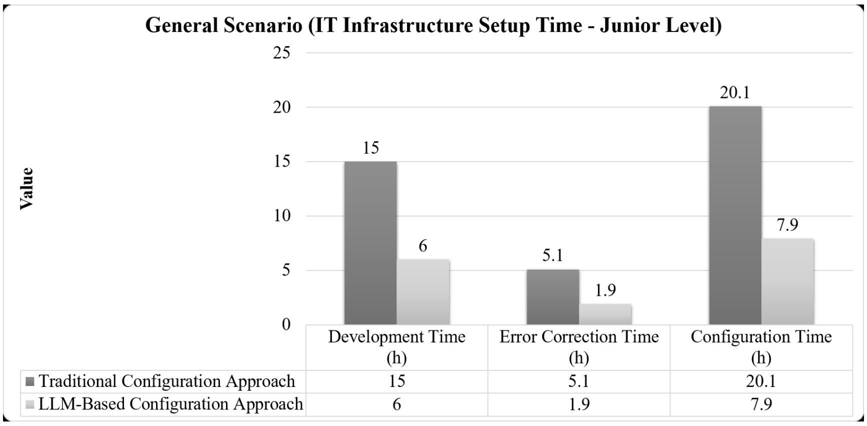


Fig. 4. Comparison of configuration time
for the generalized IT infrastructure scenario by a junior level engineer

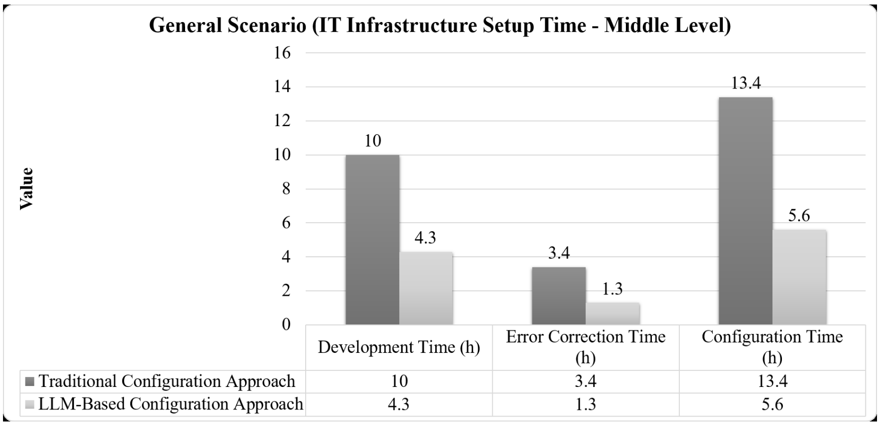


Fig. 5. Comparison of configuration time
for the generalized IT infrastructure scenario by a middle level engineer

Final results are aggregated into a generalized scenario combining data from all five cases, providing a comprehensive evaluation of the Llama 2 LLM effectiveness in infrastructure automation. The generalized scenario compared traditional and LLM-based approaches across all developer levels, highlighting overarching trends and automation advantages [6].

Analysis of development and configuration time

Research on language model scaling [38] and computational resource optimization [39] indicates that reduced computational costs and enhanced model performance indirectly shorten infrastructure development and configuration time. Our findings align with these insights.

For the generalized scenario, the traditional approach required 15 hours for junior developers, whereas the proposed method reduced this to 6 hours (Fig. 4). Similar trends are observed for middle level engineers (10 to 4.3 hours; Fig. 5) and senior engineers (8 to 3.4 hours; Fig. 6). These results demonstrate a 57–60% reduction in configuration time across all expertise levels.

Error reduction

The LLM-based approach significantly reduced error rates across all development stages [40]. Under the traditional method, the total error rate for the generalized scenario was 25%, which dropped to 8% with the proposed method (Fig. 7). Errors decreased at every stage: compilation errors fall from 9%

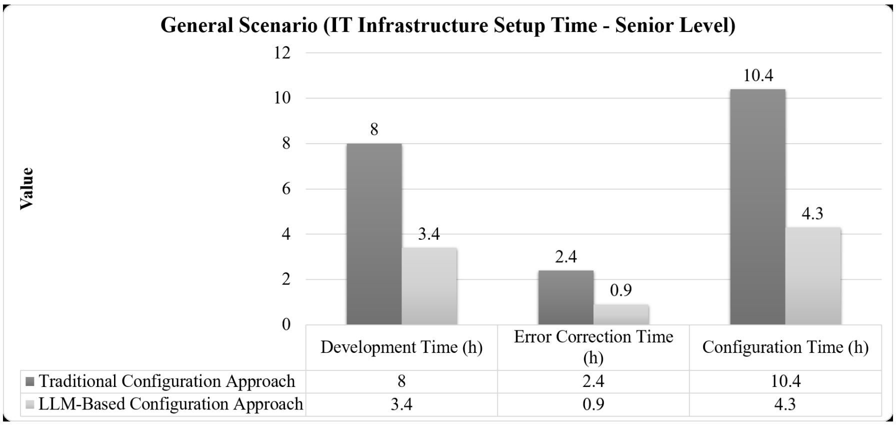


Fig. 6. Comparison of configuration time
for the generalized IT infrastructure scenario by a senior level engineer

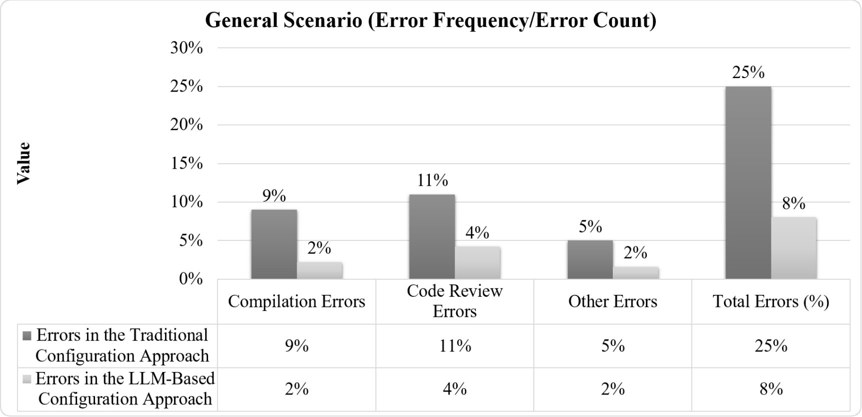


Fig. 7. Comparison of error rates in the generalized IT infrastructure configuration scenario

to 2% and code review errors from 11% to 4%. This confirms that LLM-driven automation enhances reliability by minimizing human error.

Solution quality and efficiency

For the generalized scenario, solution quality (measured as LOC to error ratio) improved by an average factor of 3.2 with the LLM-based method (Fig. 8). Calculation of the solution quality is made by the formula:

$$Q = \frac{A_{code}}{A_{error}},$$

where A_{code} is the LOC; A_{error} is the error ratio.

Solution efficiency, defined as the ratio of LOC to average configuration time, was 2.5 times higher compared to the traditional approach (Fig. 9). Calculation of the solution efficiency is made by the formula:

$$E = \frac{A_{code}}{T_{average}},$$

where A_{code} is the LOC; $T_{average}$ – average configuration time.

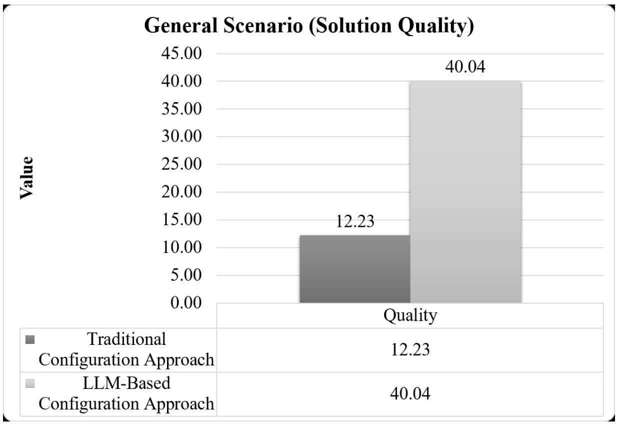


Fig. 8. Comparison of quality metrics for the generalized IT infrastructure scenario: traditional vs LLM-based approach

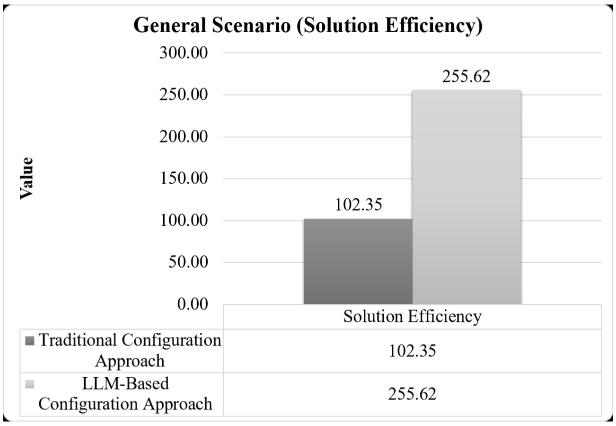


Fig. 9. Comparison of efficiency metrics for the generalized IT infrastructure scenario: traditional vs LLM-based approach

Thus, the results obtained for the generalized scenario demonstrate that the LLM not only reduces configuration time and lowers error rates, but also enhances overall solution quality and efficiency across all developer expertise levels.

General conclusions

The experiment demonstrates that the LLM-based approach reduces development time by up to 60%, lowers error rates from 25% to 8% and improves solution quality and efficiency by factors of 3.1 and 2.5, respectively. These results validate the hypothesis that LLMs like Llama 2 can effectively automate infrastructure configuration, particularly in complex, large-scale IT projects. The method offers substantial time and resource savings while enhancing solution reliability and quality compared to traditional approaches [41].

However, there are limitations for suggested approach usage. They are listed below.

1. *Effectiveness in low-complexity tasks*. In scenarios requiring the installation and configuration of a single component, traditional approaches utilizing pre-engineered scripts demonstrate superior efficiency. This is attributed to the overhead of time spent configuring AI agents when optimized pre-existing solutions tailored to specific infrastructural conditions are already available.

2. *Resource intensity of infrastructure*. To ensure acceptable query processing speeds, a cloud infrastructure supporting LLMs is required, including the allocation of GPU-accelerated instances. This

requirement introduces significant operational expenditures (OPEX) associated with leasing and maintaining computational resources.

3. *Uncertainty in outcomes for complex queries*. When generating configurations for high-level or multi-component tasks, non-deterministic outputs may arise, necessitating mandatory verification and manual changes (if required) by engineers. This limitation reduces system autonomy and increases overall deployment time.

Conclusion

This study proposes a method for automating project IT infrastructure configuration, leveraging the LLM Llama 2 to convert human-readable task descriptions in natural language into executable commands. The approach reduces the time required for IT infrastructure setup by up to 60% compared to traditional tool-based methods. A software architecture implementing the proposed method is developed in the form of an AI agent, which has demonstrated its practical efficiency.

Experimental results revealed that automation via the proposed method significantly reduces the error rate in generated software configurations. Specifically, the traditional manual configuration approach resulted in an error rate of 25%, whereas the proposed method reduced this figure to 8%. These findings highlight the substantial advantages of the proposed method over conventional tool-based configuration techniques.

Further experimental evaluations quantified the quality of automated configuration using Llama 2, demonstrating an improvement in 3 times on average compared to traditional tools. Additionally, the efficiency of the solution increased by a factor of 2.5.

Future research directions include integrating LLMs [20] with other automation tools such as configuration management systems (e.g., Ansible, Terraform) and container orchestration platforms (e.g. Kubernetes). It is also critical to explore the potential of fine-tuning the models for domain-specific tasks [42], which could enhance their accuracy and adaptability.

REFERENCES

1. **Ivlev V.A., Nikiforov I.V., Yusupova O.A.** Automation method for configuring IT infrastructure for IT projects. *International Conference on Digital Transformation: Informatics, Economics, and Education* (*DTIEE2023*), 2023, Vol. 12637, Pp. 67–73. DOI: 10.1117/12.2680779

2. **Ustinova V.E., Lutsenko A.S., Shpak A.V. et al.** A method for finding the correspondence between a railway station model and its visual representation based on graphs. *Computing, Telecommunications and Control*, 2024, Vol. 17, No. 4, Pp. 64–77. DOI: 10.18721/JCSTCS.17406

3. **Vijayakumar K., Arun C.** Automated risk identification using NLP in cloud based development environments. *Journal of Ambient Intelligence and Humanized Computing*, 2017, Pp. 1–13. DOI: 10.1007/s12652-017-0503-7

4. **Anil R. et al.** Palm 2 technical report. *arXiv:2305.10403*, 2023. DOI: 10.48550/arXiv.2305.10403

5. **Ivlev V.A., Mironenkov G.V., Nikiforov I.V., Ustinov S.M.** Ispol'zovanie modeli GPT-3 dlia generatsii informatsionno-tekhnologicheskoi infrastruktury proekta na osnove neformalizovannykh trebovanii [Using the GPT-3 model to generate a project's information technology infrastructure based on informal requirements]. *Sovremennye tekhnologii v teorii i praktike programmirovaniia* [*Modern technologies in the theory and practice of programming*], 2024, Pp. 174–176.

6. **Allamanis M., Barr E.T., Devanbu P., Sutton C.** A survey of machine learning for big code and naturalness. *arXiv:1709.06182*, 2017. DOI: 10.48550/arXiv.1709.06182

7. **Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C.L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J.,**

Lowe R. Training language models to follow instructions with human feedback. *arXiv:2203.02155*, 2022. DOI: 10.48550/arXiv.2203.02155

8. **Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L.** BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv:1910.13461*, 2019. DOI: 10.48550/arXiv.1910.13461

9. **Srivastava A. et al.** Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv:2206.04615*, 2022. DOI: 10.48550/arXiv.2206.04615

10. **Ziegler D.M., Stiennon N., Wu J., Brown T.B., Radford A., Amodei D., Christiano P., Irving G.** Fine-tuning language models from human preferences. *arXiv:1909.08593*, 2019. DOI: 10.48550/arXiv.1909.08593

11. **Chen M. et al.** Evaluating large language models trained on code. *arXiv:2107.03374*, 2021. DOI: 10.48550/arXiv:2107.03374

12. **Feng Z., Guo D., Tang D., Duan N., Feng X., Gong M., Shou L., Qin B., Liu T., Jiang D., Zhou M.** CodeBERT: A pre-trained model for programming and natural languages. *arXiv:2002.08155*, 2020. DOI: 10.48550/arXiv:2002.08155

13. **Black S., Biderman S., Hallahan E., Anthony Q., Gao L., Golding L., He H., Leahy C., McDonell K., Phang J., Pieler M., Sai Prashanth USVSN, Purohit S., Reynolds L., Tow J., Wang B., Weinbach S.** GPT-NeoX-20B: An open-source autoregressive language model. *arXiv:2204.06745*, 2022. DOI: 10.48550/arXiv:2204.06745

14. **Ivlev V.A., Mironenkov G.V., Nikiforov I.V.** Sozdanie infrastruktury proekta s pomoshch'iu neironnoi seti [Creating a project infrastructure using a neural network]. *Nedelia nauki IKNK* [*Institute of Computer Science and Cybersecurity Science Week*], 2024. Pp. 24–26.

15. **Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P.J.** Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683*, 2019. DOI: 10.48550/arXiv:1910.10683

16. **Schick T., Dwivedi-Yu J., Dessì R., Raileanu R., Lomeli M., Zettlemoyer L., Cancedda N., Scialom T.** Toolformer: Language models can teach themselves to use tools. *arXiv:2302.04761*, 2023. DOI: 10.48550/arXiv:2302.04761

17. **Gururangan S., Marasović A., Swayamdipta S., Lo K., Beltagy I., Downey D., Smith N.A.** Don't stop pretraining: Adapt language models to domains and tasks. *arXiv:2004.10964*, 2020. DOI: 10.48550/arXiv:2004.10964

18. **Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., von Platen P., Ma C., Jernite Y., Plu J., Xu C., Le Scao T., Gugger S., Drame M., Lhoest Q., Rush A.** Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, Pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6

19. **Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.** Attention is all you need. *arXiv:1706.03762*, 2017. DOI: 10.48550/arXiv.1706.03762

20. **Chowdhery A. et al.** PaLM: Scaling language modeling with pathways. *arXiv:2204.02311*, 2022. DOI: 10.48550/arXiv.2204.02311

21. **Touvron H. et al.** Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023. DOI: 10.48550/arXiv:2307.09288

22. **Brown T.B. et al.** Language models are few-shot learners. *arXiv:2005.14165*, 2020. DOI: 10.48550/arXiv:2005.14165

23. **Gao L., Biderman S., Black S., Golding L., Hoppe T., Foster C., Phang J., He H., Thite A., Nabeshima N., Presser S., Leahy C.** The Pile: An 800GB dataset of diverse text for language modeling. *arXiv:2101.00027*, 2020. DOI: 10.48550/arXiv:2101.00027

24. **Le Scao T. et al.** BLOOM: A 176B-parameter open-access multilingual language model *arXiv:2211.05100*, 2022. DOI: 10.48550/arXiv:2211.05100

25. **Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E., Le Q., Zhou D.** Chain-of-thought prompting elicits reasoning in large language models. *arXiv:2201.11903*, 2022. DOI: 10.48550/arXiv:2201.11903

26. **Black S., Leo G., Wang P., Leahy C., Biderman S.** GPT-neo: Large scale autoregressive language modeling with mesh-tensorflow (1.0). *Zenodo*, 2021. DOI: 10.5281/zenodo.5297715.

27. **Bengio Y., Louradour J., Collobert R., Weston J.** Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, Pp. 41–48.

28. **Rozière B. et al.** Code Llama: Open foundation models for code. *arXiv:2308.12950*, 2023. DOI: 10.48550/arXiv:2308.12950

29. **Rajani N.F., McCann B., Xiong C., Socher R.** Explain yourself! Leveraging language models for commonsense reasoning. *arXiv:1906.02361*, 2019. DOI: 10.48550/arXiv:1906.02361

30. **Armitage J., Kacupaj E., Tahmasebzadeh G., Swati, Maleshkova M., Ewerth R., Lehmann J.** MLM: A benchmark dataset for multitask learning with multiple languages and modalities. *arXiv:2008.06376*, 2020. DOI: 10.48550/arXiv:2008.06376

31. **Devlin J., Chang M.-W., Lee K., Toutanova K.** BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018. DOI: 10.48550/arXiv:1810.04805

32. **Ivlev V.A., Mironenkov G.V., Nikiforov I.V., Kovalev A.D.** Generatsiia informatsionno-tekhnologicheskoi infrastruktury proekta na osnove neformalizovannykh trebovanii [Generation of the project's information technology infrastructure based on informal requirements]. *Sovremennye tekhnologii v teorii i praktike programmirovaniia* [*Modern technologies in the theory and practice of programming*], 2023, Pp. 242–244.

33. **Bommasani R. et al.** On the opportunities and risks of foundation models. *arXiv:2108.07258*, 2021. DOI: 10.48550/arXiv:2108.07258

34. **Nikiforov I.V., IUsupova O.A., Voinov N.V., Kovalev A.D., Tkachuk A.S., Varlamov D.A., Geras'kin E.V.** *Programmnye instrumenty obrabotki i vizualizatsii dannykh. Elasticsearch, Logstash, Kibana, Grafana, Prometheus* [*Software tools for data processing and visualization. Elasticsearch, Logstash, Kibana, Grafana, Prometheus*]. St. Petersburg: POLITEKH-PRESS, 2023. DOI: 10.18720/SPBPU/2/id23-74

35. **Fried D., Aghajanyan A., Lin J., Wang S., Wallace E., Shi F., Zhong R., Yih W.-t., Zettlemoyer L., Lewis M.** InCoder: A generative model for code infilling and synthesis. *arXiv:2204.05999*, 2022. DOI: 10.48550/arXiv:2204.05999

36. **Nijkamp E., Pang B., Hayashi H., Tu L., Wang H., Zhou Y., Savarese S., Xiong C.** CodeGen: An open large language model for code with multi-turn program synthesis. *arXiv:2203.13474*, 2022. DOI: 10.48550/arXiv:2203.13474

37. **Sajja P.S.** Computer-assisted tools for software development. In: *Essence of Systems Analysis and Design: A Workbook Approach*, 2017, Pp. 93–105. DOI: 10.1007/978-981-10-5128-9_5

38. **Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., Amodei D.** Scaling laws for neural language models. *arXiv:2001.08361*, 2020. DOI: 10.48550/arXiv:2001.08361

39. **Hoffmann J. et al.** Training compute-optimal large language models. *arXiv:2203.15556*, 2022. DOI: 10.48550/arXiv:2203.15556

40. **Hindle A., Barr E., Gabel M., Su Z., Devanbu P.** On the naturalness of software. *Communications of the ACM*, 2016, Vol. 59, No. 5, Pp. 122–131.

41. **Lu S. et al.** CodeXGLUE: A machine learning benchmark dataset for code understanding and generation. *arXiv:2102.04664*, 2021. DOI: 10.48550/arXiv:2102.04664

42. **Bahdanau D., Cho K., Bengio Y.** Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014. DOI: 10.48550/arXiv:1409.0473

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Ivlev Vladislav A.
Ивлев Владислав Александрович
E-mail: nevidd@yandex.ru

Nikiforov Igor V.
Никифоров Игорь Валерьевич
E-mail: igor.nikiforovv@gmail.com
ORCID: https://orcid.org/0000-0003-0198-1886

Ustinov Sergey M.
Устинов Сергей Михайлович
E-mail: usm50@yandex.ru
ORCID: https://orcid.org/0000-0003-4088-4798

Submitted: 13.03.2025; Approved: 13.05.2025; Accepted: 26.05.2025.

Поступила: 13.03.2025; Одобрена: 13.05.2025; Принята: 26.05.2025.

Circuits and Systems for Receiving, Transmitting and Signal Processing

Устройства и системы передачи, приема и обработки сигналов

Short message
DOI: https://doi.org/10.18721/JCSTCS.18207
UDC 621.3.049.774.2

INCREMENTAL DELTA-SIGMA MODULATOR

M.M. Pilipko ✉, *D.V. Morozov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ m_m_pilipko@rambler.ru

Abstract. A delta-sigma modulator with reset for incremental ΔΣ ADCs for the 180 nm CMOS technology with a supply voltage of 3.3 V from Mikron JSC is presented. The simulation of the ΔΣ modulator in the time domain in the Virtuoso analog design environment from Cadence DS was performed. The clock frequency was set to 6.25 MHz. The power consumption was about 9.5 mW. The reset was performed every 32 or 128 clock cycles. The results of the ΔΣ modulator simulation were processed in MATLAB. The digital decimation filter in the form of a cascade of integrators was realized in software. At the oversampling ratio of 32, the modulator shows SINAD = 69.3 dB (ENOB = 11.2 bits) and SFDR = 76.9 dB. At the oversampling ratio of 128, SINAD = 88.7 dB (ENOB = 14.4 bits) and SFDR = 92.7 dB are achieved. The crystal dimensions were 640 x 340 μm. The ΔΣ modulator circuit is suitable for precise digitization of sensor signals in the audio frequency range.

Keywords: analog-to-digital converter, delta-sigma modulator, incremental delta-sigma ADC, bootstrapped switch, dynamic element matching

Acknowledgements: The production of the integrated microcircuit was funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of the federal project “Training of personnel and scientific foundation for the electronic industry” under the state assignment for the implementation of the research work “Development of a methodology for prototyping an electronic component base in domestic microelectronic production based on the MPW service”.

Citation: Pilipko M.M., Morozov D.V. Incremental delta-sigma modulator. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 91–98. DOI: 10.18721/JCSTCS.18207

Краткое сообщение
DOI: https://doi.org/10.18721/JCSTCS.18207
УДК 621.3.049.774.2

ИНКРЕМЕНТАЛЬНЫЙ ДЕЛЬТА-СИГМА МОДУЛЯТОР

М.М. Пилипко ✉, *Д.В. Морозов*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ m_m_pilipko@rambler.ru

Аннотация. Представлен ΔΣ модулятор со сбросом для инкрементальных ΔΣ АЦП по 180 нм КМОП-технологии с напряжением питания 3,3 В от АО «Микрон». Моделирование ΔΣ модулятора во временной области проводилось в среде аналогового проектирования Virtuoso компании Cadence DS. Тактовая частота была равна 6,25 МГц. Потребляемая мощность составила около 9,5 мВт. Сброс производился каждые 32 или 128 тактов. Результаты моделирования ΔΣ-модулятора обрабатывались в MATLAB. Цифровой децимирующий фильтр в виде каскада интеграторов реализован программно. При коэффициенте передискретизации 32 модулятор обеспечивает SINAD = 69,3 дБ (ENOB = 11,2 бит) и SFDR = 76,9 дБ. При коэффициенте передискретизации 128 достигаются SINAD = = 88,7 дБ (ENOB = 14,4 бит) и SFDR = 92,7 дБ. Размеры кристалла составили 640x340 мкм. Схема ΔΣ модулятора пригодна для точной оцифровки сигналов датчиков физических величин в звуковом диапазоне частот.

Ключевые слова: аналого-цифровой преобразователь, дельта-сигма модулятор, инкрементальный дельта-сигма АЦП, ключ с постоянным сопротивлением, динамическое согласование элементов

Финансирование: Производство интегральной микросхемы было выполнено за счет средств Министерства науки и высшего образования Российской Федерации в рамках федерального проекта «Подготовка кадров и научного фундамента для электронной промышленности» по государственному заданию на выполнение научно-исследовательской работы «Разработка методики прототипирования электронной компонентной базы на отечественных микроэлектронных производствах на основе сервиса MPW».

Для цитирования: Pilipko M.M., Morozov D.V. Incremental delta-sigma modulator // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 91–98. DOI: 10.18721/JCSTCS.18207

Introduction

For battery-powered systems, such as sensors, power-efficient analog-to-digital converters (ADCs) are especially important and are typically tuned to the required bandwidth. Incremental delta-sigma (ΔΣ) ADCs [1, 2] are the optimal choice for achieving high power efficiency. Incremental ΔΣ ADCs have a reset and perform a sample-by-sample conversion, which distinguishes them from traditional ΔΣ ADCs, when digitizing weakly correlated input samples. Since such converters typically have a finite impulse response, the corresponding decimation filter can be as simple as a cascade of integrators, which is much simpler than their counterparts using filters with an infinite impulse response.

The block diagram of such an ADC is shown in Fig. 1. It consists of a ΔΣ modulator, a decimation filter and a reset circuit. The input signal X is fed to the ΔΣ modulator. At the ΔΣ modulator output, an oversampled digital data stream Y is formed, which is processed by the decimation filter. At the output of the digital decimation filter, the ΔΣ ADC output code D is formed. The clock signal (clk) is fed both to the ΔΣ modulator and the reset circuit. The reset circuit is a counter that is controlled by an oversampling ratio signal (OSR) – 32 or 128 in this paper – and discretely changes the conversion

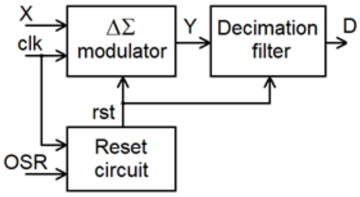


Fig. 1. Incremental ΔΣ ADC

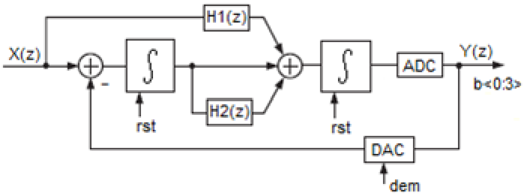


Fig. 2. The second-order ΔΣ modulator with reset

factor within the oversampling ratio values and generates a reset signal (rst). This paper will focus on the design of the ΔΣ modulator with the reset for incremental ΔΣ ADCs.

Functional-level model of the ΔΣ modulator

Fig. 2 shows a functional-level model in z-domain [3, 4] of the second-order ΔΣ modulator with reset, which consists of two integrators with reset, feedforward paths $H1(z) = H2(z) = 1–z^{-1}$, two analog adders, a quantizing circuit in the form of an ADC, and a digital-to-analog converter (DAC) in the negative feedback loop. Unlike other modulator structures, this structure requires only one feedback DAC and does not require an adder before the local ADC. The input is designated as X(z). The digital data stream b<0:3> is formed at the output Y(z) of the ΔΣ modulator. The first integrator has a transfer function $z^{-1}/(1–z^{-1})$, and the second integrator has a transfer function $0.5/(1–z^{-1})$. Both integrators are reset by the "rst" signal every 32 or 128 clock cycles. The feedback DAC suffers from mismatch of component values. In order to alleviate this problem, a dynamic elements matching (DEM) circuit in DAC is developed [5]. The signal for controlling the DEM part of the DAC is designated as "dem".

Cadence Virtuoso circuit of the ΔΣ modulator

According to the functional level block diagram of the ΔΣ modulator shown in Fig. 2, *a* circuit of the ΔΣ modulator based on switched capacitors has been developed in the Virtuoso analog design environment from Cadence DS. The circuit of the ΔΣ modulator with reset is shown in Fig. 3, where the 180 nm complementary metal-oxide-semiconductor (CMOS) technology with a supply voltage of 3.3 V from Mikron JSC was used. The ΔΣ modulator layout is shown in Fig. 4. Sizes of the layout are 640×340 um.

In Fig. 3, input signals – non-inverting (inp) and inverting (inm) – are applied to the integrators via bootstrapped switches [6, 7]. Each integrator utilizes a folded-cascode rail-to-rail operational transconductance amplifier (OTA) similar to [5, 8], with the width of the transistors in OTA1 being two times greater than the width of the transistors in OTA2 due to a proportionally larger load. The OTA inputs are non-inverting (vp) and inverting (vm), the OTA outputs are non-inverting (vop) and inverting (vom). The signal equal to half the supply voltage is designated as "vcm". The output signals of the integrators are "o1m" and "o1p" for the first integrator and "o2m" and "o2p" for the second integrator.

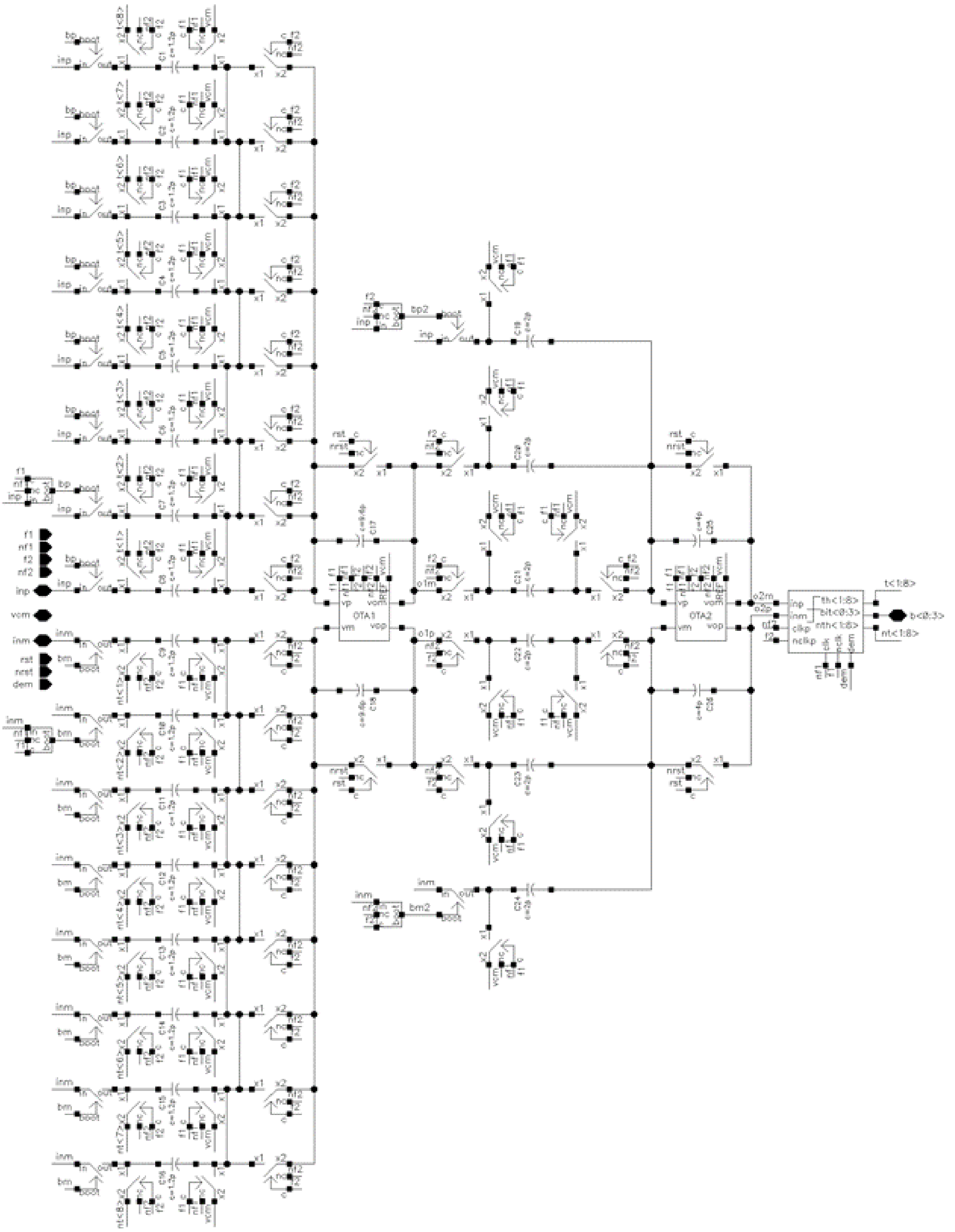


Fig. 3. The ΔΣ modulator in Virtuoso by Cadence DS

The quantizer at the output of the ΔΣ modulator forms four output bits b<0:3> and has eight quantization levels, which are provided by comparators similar to [5]. The non-inverting and inverting feedback signals of the ΔΣ modulator are presented in thermometer code and are designated by t<1:8> and nt<1:8>, respectively. The signal for controlling the dynamic elements matching circuit [5] is designated

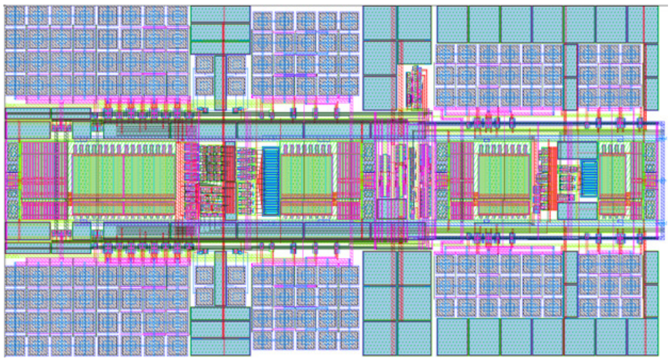


Fig. 4. The ΔΣ modulator layout

as "dem". The circuit uses CMOS switches controlled by two phase sequences "f1" and "f2" similar to [5]. Phase sequences "nf1" and "nf2" are inverted to "f1" and "f2", respectively. The main feature of this ΔΣ modulator circuit is the presence of CMOS switches that perform reset of the integrators by connecting the input and the output of the OTA pairwise using the "rst" signal ("nrst" is the inverse signal to "rst").

The input capacitance of the ΔΣ modulator consists of eight parallel-connected capacitors with a nominal value of 1.2 pF to both non-inverting (inp) and the inverting (inm) inputs, which totally gives 9.6 pF in both of the specified nodes. Capacitors with a nominal value of 9.6 pF are connected in the feedback loops of OTA1. This ratio of capacitances sets the unity gain in the first integrator of the ΔΣ modulator. The load of the first integrator is the input capacitance of the second integrator, which in total is 4 pF at both the non-inverting (vop) and inverting (vom) outputs of OTA1. Capacitors with a nominal value of 4 pF are connected in the feedback loops of OTA2. Each of the capacitors connected to both the non-inverting (vp) and the inverting (vm) inputs of OTA2 has a nominal value of 2 pF. This ratio of capacitances sets the gain of 1/2 in the second integrator of the ΔΣ modulator. With these ratios of capacitances, the circuit of the ΔΣ modulator corresponds to its z-domain transfer function in the functional-level model described in previous section.

The ΔΣ modulator circuit in Virtuoso by Cadence DS was simulated at 27°C in the time domain with transient noise option turned on. The clock frequency was set to 6.25 MHz. The input harmonic signal frequency was 1335 Hz, the amplitude was 1 V. Power consumption was 9.5 mW. The simulation results were processed in MATLAB. The input differential signal "inp" and "inm", reset signal (rst) and output code b<0:3> in decimal form (DSM code) are shown in Fig. 5 for the case when the "rst" signal acts every 32 clock cycles. The ΔΣ modulator output code in decimal form takes values from 0 to 8. After the reset pulse on the second graph, the code at the output of the ΔΣ modulator takes the value 4, which corresponds to the middle of the specified range of the output values.

Processing the simulation results of the ΔΣ modulator in MATLAB

The results of the time-domain simulation of the ΔΣ modulator were processed in MATLAB. A software implementation was done for the decimation filter from Fig. 1 in the form of a cascade of integrators. The number of the integrators in the decimating filter is two, which corresponds to the order of the ΔΣ modulator.

For the case when the reset signal (rst) acts every 32 clock cycles, the output code after processing by the digital decimation filter is shown in Fig. 6, *a*. Every 32 clock cycles, the sum of the ΔΣ modulator output codes is formed by accumulation. The resulting digital values before the reset moment proportionally represent the shape of the input harmonic analog signal. Fig. 6, *b* shows the decimated output code of the digital filter fixed in the register before the reset moment. In decimal form, the

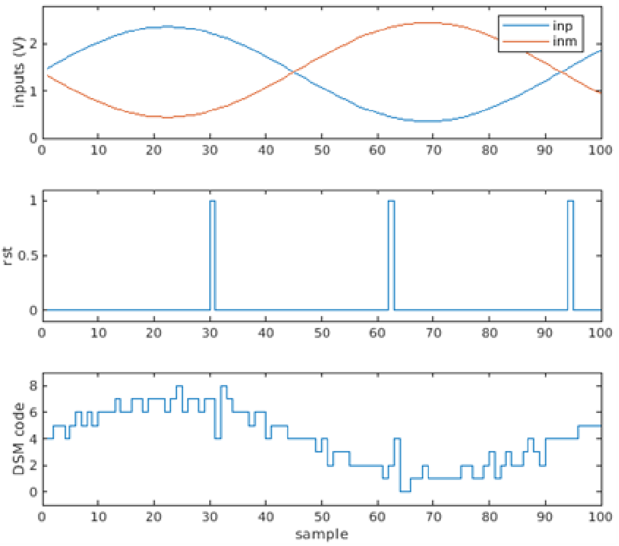


Fig. 5. Simulation results of the ΔΣ modulator from Virtuoso by Cadence DS processed in MATLAB

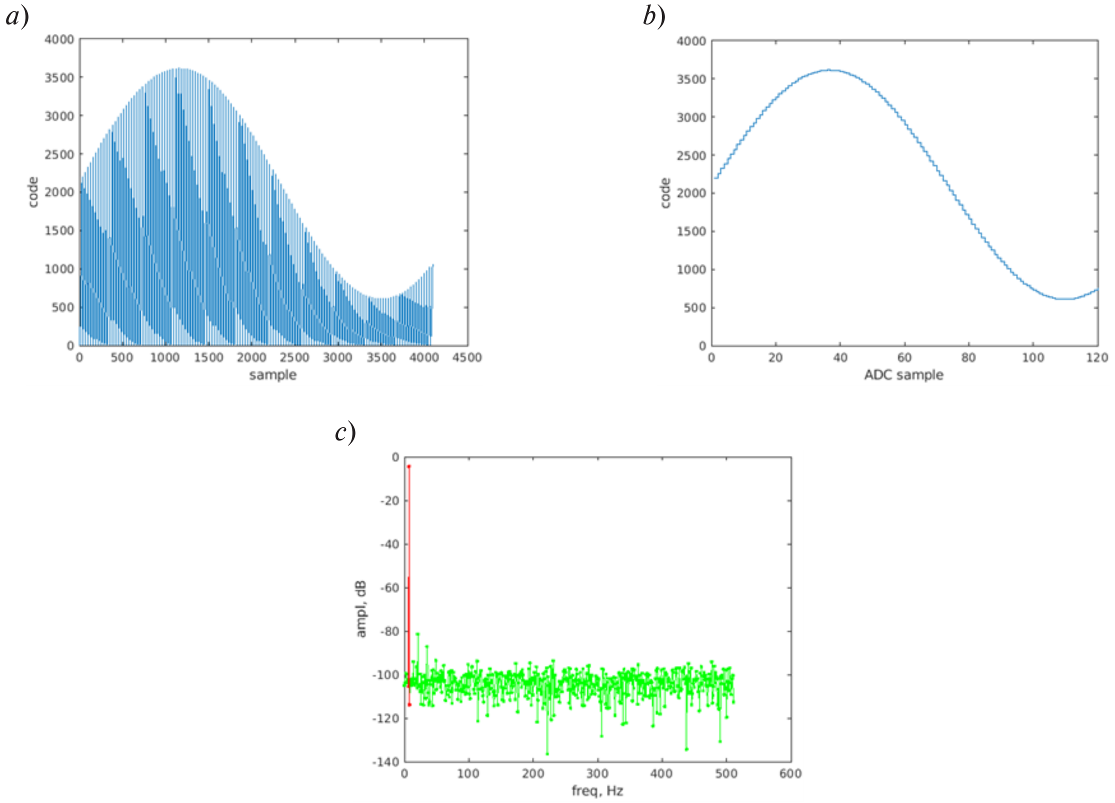


Fig. 6. Processing in MATLAB the simulation results for the reset signal (rst) acting every 32 clock cycles

values of the decimation filter output code range from 608 to 3616. The decimation filter output code spectrum is shown in Fig. 6, *c*, a 1024-point discrete Fourier transform with a rectangular window was performed. The calculated signal-to-noise and distortion ratio is SINAD = 69.3 dB (the effective number of bits is ENOB = 11.2 bits), spurious-free dynamic range is SFDR = 76.9 dB.

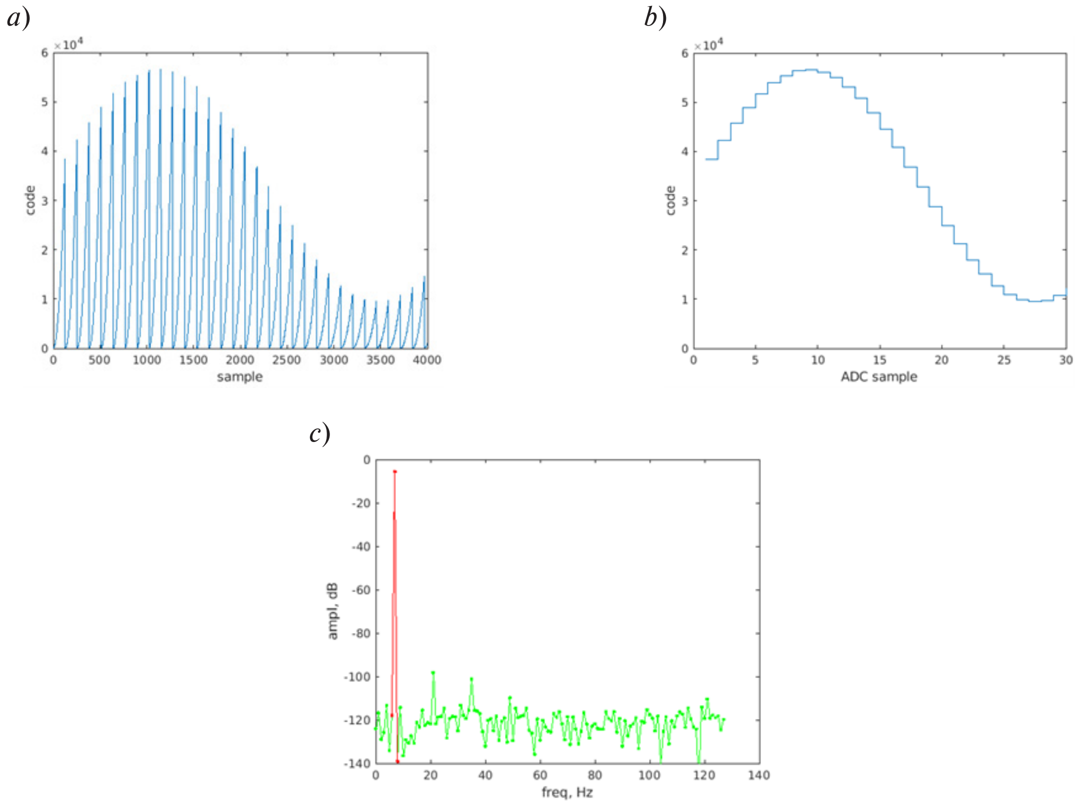


Fig. 7. Processing in MATLAB the simulation results for the reset signal (rst) acting every 128 clock cycles

For the case when the reset signal (rst) acts every 128 clock cycles, the output code after processing by the decimation filter in the form of the cascade of integrators is shown in Fig. 7, *a*. Every 128 clock cycles, the sum of the ΔΣ modulator output codes is formed by accumulation. The clock frequency and the input harmonic signal are the same as in the previous case.

Fig. 7, *b* shows the output code of the decimation filter fixed in the register before the reset moment. In decimal form, the values of the decimation filter output code range from 9466 to 56584. The decimation filter output code spectrum is shown in Fig. 7, *c*, a 256-point discrete Fourier transform with a rectangular window was performed. The calculated characteristics are SINAD = 88.7 dB (ENOB = 14.4 bits) and SFDR = 92.7 dB.

Conclusions

Delta-sigma modulators are suitable for relatively low-frequency applications, such as sensor systems and audio applications, that require high-quality digitization of the input signal. Unlike traditional ΔΣ ADCs, incremental ΔΣ ADCs allow easy integration into multi-channel systems due to sample-by-sample digitization without memory effect.

The ΔΣ modulator with reset for incremental ΔΣ ADCs was designed in 180 nm CMOS technology with a supply voltage of 3.3 V from Mikron JSC. The clock frequency is set to 6.25 MHz. Power consumption is about 9.5 mW. The reset acted every 32 or 128 clock cycles, i.e., the signal band was from 0 to either 195 kHz or 49 kHz, respectively. When the reset acts every 32 clock cycles, circuit properties are as follows: SINAD = 69.3 dB (ENOB = 11.2 bits), SFDR = 76.9 dB. With the reset at every 128 clock cycles, SINAD = 88.7 dB (ENOB = 14.4 bits) and SFDR = 92.7 dB. This allows the ΔΣ modulator to be used in precision measurement systems.

REFERENCES

1. **Tan Z., Chen C.-H., Chae Y., Temes G.C.** Incremental delta-sigma ADCs: A tutorial review. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2020, Vol. 67, No. 12, Pp. 4161–4173. DOI: 10.1109/TCSI.2020.3033458

2. **Satyshev V.I.** Modern approaches to design of multi-channel delta-sigma ADCs. *Computing, Telecommunications and Control*, 2022, Vol. 15, No. 2, Pp. 25–31. DOI: 10.18721/JCSTCS.15202

3. **San H., Konagaya H., Xu F., Motozawa A., Kobayashi H., Ando K.** Second-order ΔΣAD modulator with novel feedforward architecture. *50th Midwest Symposium on Circuits and Systems*, 2007, Pp. 148–151. DOI: 10.1109/MWSCAS.2007.4488558

4. **Honarparvar M., de la Rosa J.M., Sawan M.** A 0.9-V 100-μ W feedforward adder-less inverter-based MASH ΔΣ modulator with 91-dB dynamic range and 20-kHz bandwidth. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2018, Vol. 65, No. 11, Pp. 3675–3687. DOI: 10.1109/TCSI.2018.2854220

5. **Pilipko M.M., Morozov D.V., Yenuchenko M.S.** MASH 2-2 delta-sigma modulator with dynamic element matching in 0.18 μm CMOS technology. *Computing, Telecommunications and Control*, 2023, Vol. 16, No. 3, Pp. 29–38. DOI: 10.18721/JCSTCS.16303

6. **Razavi B.** The bootstrapped switch [A circuit for all seasons]. *IEEE Solid-State Circuits Magazine*, 2015, Vol. 7, No. 3, Pp. 12–15. DOI: 10.1109/MSSC.2015.2449714

7. **Kim S.-H., Lee Y.-H., Chung H.-J., Jang Y.-C.** A bootstrapped analog switch with constant on-resistance. *IEICE Transactions on Electronics*, 2011, Vol. E94.C, No. 6, Pp. 1069–1071. DOI: 10.1587/transele.E94.C.1069

8. **Pilipko M.M., Morozov D.V., Yenuchenko M.S.** Delta-sigma modulator with 10 MHz clock frequency in 180 nm CMOS technology. *MES-2018*, 2018, Pp. 44–48. DOI: 10.31114/2078-7707-2018-4-44-48

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Pilipko Mikhail M.
Пилипко Михаил Михайлович
E-mail: m_m_pilipko@rambler.ru
ORCID: https://orcid.org/0000-0003-3813-6846

Morozov Dmitry V.
Морозов Дмитрий Валерьевич
E-mail: dvmorozov@inbox.ru
ORCID: https://orcid.org/0000-0003-3403-0120

Submitted: 17.03.2025; Approved: 24.05.2025; Accepted: 30.05.2025.
Поступила: 17.03.2025; Одобрена: 24.05.2025; Принята: 30.05.2025.

Short message
DOI: https://doi.org/10.18721/JCSTCS.18208
UDC 621.3.049.77

STABILIZED REFERENCE CURRENT SOURCE FOR BIOMEDICAL APPLICATIONS

K.A. Mironov ✉, *D.V. Morozov*, *D.B. Akhmetov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ mironov.ka@edu.spbstu.ru

Abstract. Neurostimulators are devices used to electrically stimulate the nervous system. They are a promising alternative to existing pharmacological methods of treating neurological disorders. The paper presents the current driver, one of the key blocks for providing electrical stimulation. The basic requirements and characteristics of this device are described. The noise of the reference current source, current mirror has been analyzed and the effect of the differential amplifier noise on the total noise current at the output devices has been considered. Based on the results of the analysis, a method for estimating the required output impedance of the current driver is proposed. The circuit implemented in 180 nm CMOS process. The output impedance of not less than 30 MOhm is obtained at the output current of 140 μA and the voltage compliance of 90.9% of the supply voltage. A comparative analysis of the results with the work of other authors is given.

Keywords: current source, neurostimulation, noise PSD, output impedance, feedback

Acknowledgements: The production of the integrated microcircuit was funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of the federal project “Training of personnel and scientific foundation for the electronic industry” under the state assignment for the implementation of the research work “Development of a methodology for prototyping an electronic component base in domestic microelectronic production based on the MPW service”.

Citation: Mironov K.A., Morozov D.V., Akhmetov D.B. Stabilized reference current source for biomedical applications. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 99–110. DOI: 10.18721/JCSTCS.18208

Краткое сообщение
DOI: https://doi.org/10.18721/JCSTCS.18208
УДК 621.3.049.77

СТАБИЛИЗИРОВАННЫЙ ИСТОЧНИК ТОКА ДЛЯ БИОМЕДИЦИНСКИХ ПРИМЕНЕНИЙ

К.А. Миронов ✉, *Д.В. Морозов*, *Д.Б. Ахметов*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ mironov.ka@edu.spbstu.ru

Аннотация. Нейростимуляторы являются многообещающей альтернативой существующим фармацевтическим методам при лечении широкого спектра неврологических заболеваний. В работе рассмотрена реализация драйвера тока, применяемого в нейростимуляторах. Описаны основные требования и характеристики данного устройства. Предложен алгоритм расчета минимально необходимого выходного сопротивления, построенный на основе анализа шума схемы источника опорного тока. Результаты подтверждены моделированием с использованием отечественной КМОП технологии с разрешением 180 нм. Получено выходное сопротивление не менее 30 МОм при выходном токе 140 мкА и диапазоне рабочих напряжений 90,9% от напряжения питания. Приведен сравнительный анализ результатов с работами других авторов.

Ключевые слова: источник тока, нейростимуляция, спектральная плотность мощности шума, выходное сопротивление, обратная связь

Финансирование: Производство интегральной микросхемы было выполнено за счет средств Министерства науки и высшего образования Российской Федерации в рамках федерального проекта «Подготовка кадров и научного фундамента для электронной промышленности» по государственному заданию на выполнение научно-исследовательской работы «Разработка методики прототипирования электронной компонентной базы на отечественных микроэлектронных производствах на основе сервиса MPW».

Для цитирования: Mironov K.A., Morozov D.V., Akhmetov D.B. Stabilized reference current source for biomedical applications // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 99–110. DOI: 10.18721/JCSTCS.18208

Introduction

Neurological disorders, such as chronic pain, Parkinson’s disease, Alzheimer’s disease, Huntington’s disease, epilepsy, obesity and addiction, affect a large part of the world’s population. The effectiveness of currently available pharmaceutical therapies is limited, because patients become resistant to treatment with long-term use. In addition, these treatments often have unwanted side effects. An alternative treatment option is neurostimulation – changing the properties of nerve tissue through the targeted application of electrical current. The challenge is to develop neurostimulators that meet safety, energy efficiency and performance requirements.

Patient safety requires the use of current-controlled differential stimulation, in which a certain amount of charge is injected into the nerve tissue during the positive phase and is pumped out during the negative phase. One of the main safety requirements is the reduction of the residual charge, as it can cause damage to the nerve tissue or premature fibrotic tissue formation and a decrease in the patient's response. The shape of the current pulses is usually rectangular, but there are studies suggesting that pulses with complex shapes may be safer [1]. On the other hand, it is possible to generate pulses with complex shapes using a current-steering digital-to-analog converter (DAC) with appropriate control signals.

The current driver is the block that stabilizes the stimulation current amplitude and, optionally, adjusts the current amplitude. In essence, it acts as a complex current source and should therefore have the same basic requirement of high output impedance. Nevertheless, this parameter is not sufficient, and another one should be added according to the specificity of the application – the dynamic range of the output voltage or, as it is often called, the compliance voltage. The nature of this parameter is as follows. The stimulation electrodes are connected between the driver and the supply or ground level (depending on whether nMOS or pMOS current mirror is used), so the voltage drop across the electrodes should be as high as possible [1]. It can also be described as the maximum load and electrode equivalent resistance at the maximum output current, according to Ohm's law. At the lower output currents, the compliance voltage is higher due to the lower MOSFET overdrive voltage at the lower currents.

Any work devoted to the design of a stimulation unit for a neurostimulator is aimed at ensuring patient safety. Therefore, the highest possible output impedance of the stimulating unit needs to be achieved to ensure proper operation under different conditions and electrode types.

One of the traditional approaches to enhancing the output impedance of a current mirror involves extending the channel length of the current-sourcing (or current-sinking, in nMOS configurations) transistor. While this technique allows measurable improvements in output impedance, the improvement is relatively modest compared to the substantial area overhead incurred, making it unsuitable for use in implantable applications. The second method is to apply circuit techniques to stabilize the output current. One of the most obvious techniques is cascoding, in which the output impedance is proportional to the common base transistor's intrinsic gain. Although it is possible to achieve output impedances of up to 100 MΩ, there is a problem of high voltage drop across the current mirror, as discussed in [3]. A more interesting approach is to use voltage stabilization circuits based on differential or instrumental amplifiers (Op-Amp), as implemented in [1–2, 4–7]. This allows the operating point of the current sourcing transistor to be stabilized so that any variations in the output voltage are compensated by the stabilization circuit.

In addition, there is no restriction on the current sourcing transistor being in the active region, in fact the triode region can also be used to achieve higher voltage compensation. Finally, all the output impedance boosting techniques discussed can be combined. For example, in [5], an increasing channel length of transistors (up to 1um) is used and a complex stabilization circuit consisting of two 70 dB Op-Amps. The result is an output impedance of 320 GΩ. However, no one mentions the upper and lower limits on the amount of the output impedance. The lower limit should obviously be specified in the medical requirements for the device. However, a review of the literature did not reveal specific values, possibly because quantitative studies of the effect of residual charge on nerve tissue damage have not been published. In this regard, it was decided to find the output impedance limitation from above, i.e., the maximum possible current stability that we can provide with a typical circuit.

Output impedance impact on current stability

In simple terms, the stimulation process involves pumping charge into neural tissue and then compensating for the pumped charge (i.e., pumping charge in the opposite direction).

The impedance of the electrode/neural tissue interface acts as the load impedance and, to a first-order approximation, can be represented as a resistor (including the electrode and electrolyte resistances) and capacitor (including the electrode capacitance and the double layer capacitance) in series (R_L and C_L, respectively).

In deep brain stimulation (DBS) applications, maintaining precise current delivery is crucial for both therapeutic efficacy and patient safety. The load resistance in such systems typically varies between 1 and 10 kΩ [8], with these fluctuations potentially affecting both current stability and charge balance during stimulation. Understanding the nature of these impedance changes is essential for designing robust DBS systems.

Load impedance variations in DBS can be categorized into two types based on their temporal characteristics. The first are long-term impedance changes, which may be caused by formation of fibrous tissue encapsulation around implanted electrodes (a natural biological response) or by using different electrode materials or designs with varying intrinsic impedance characteristics. These occurs over several weeks or months after implantation and change slowly relative to stimulation pulse durations (typically 1 to 100 ms). Therefore, it has minimal effect on instantaneous charge balance during individual pulses, but periodic system recalibration may be required, which, however, does not significantly contribute to excess charge accumulation.

The second category includes short-term impedance fluctuations caused by dynamic polarization of neural tissue during current flow [9] or by asymmetric charge/discharge behavior of the electrode-tissue interface (modeled as a double-layer capacitor). Such fluctuations occur within each stimulation pulse phase and exhibit rapid changes on millisecond timescales. Thus, it significantly affects instantaneous current delivery and can lead to substantial charge imbalance, if not properly compensated.

The stability of the output current in the face of load variations is fundamentally determined by the output impedance $\left(R_{out}\right)$ of the current source. The relationship between load variation $\left(\Delta R_L\right)$ and resulting current variation $\left(\Delta I_{R_L}\right)$ can be analyzed using current divider principles:

$$\Delta I_{R_L} = I_{strim} \cdot \frac{\Delta R_L}{R_{out} + \Delta R_L + R_L}, \tag{1}$$

where I_{strim} is the nominal stimulation current, R_L is the baseline load resistance, ΔR_L is the impedance variation. The trade-off in implementation is that higher output impedance improves current regulation, but may compromise power efficiency. Circuit techniques like active feedback can achieve high R_{out} without excessive output voltage drop.

Enhanced current stabilization using Op-Amps in feedback circuits

One of the most effective techniques for stabilizing current in modern analog circuits involves integrating Op-Amps within a feedback loop to precisely regulate the operating points of current mirror transistors. This paper focuses on the regulated drain current mirror (RDCM) due to its superior performance in maintaining consistent output current. The output impedance of this configuration can be derived using the following equation:

$$Z_{out} \approx A \cdot g_m \cdot r_o^2, \tag{2}$$

where A is the open-loop gain of the Op-Amp, g_m is the transconductance of the MOSFET, r_o is the intrinsic output impedance of the transistor. To achieve an ultra-high output impedance exceeding 100 GΩ, the Op-Amp must provide a gain of at least 70 dB, as per the given expression. However, while the equation suggests no strict upper limit on impedance, practical constraints arise from thermal and flicker noise, which ultimately determine the maximum achievable performance.

Noise analysis and optimization strategies

The intrinsic noise of the circuit significantly impacts the stability of the output current. To assess this, the power spectral density (PSD) of the noise at the driver's output was analyzed, considering contributions from: the reference current source, the current mirror transistors, the Op-Amp.

The RDCM amplifies the reference current by a factor K, a common approach to maintaining current accuracy while minimizing power dissipation and chip area. The total output noise PSD $\left(S_{i_{out}}\right)$ is expressed as:

$$S_{i_{out}} = S_{i_{drv}} + K^2 \cdot S_{i_{ref}}, \tag{3}$$

when $K > 1$, the reference current noise becomes the dominant contributor. Using the SPICE LEVEL 2 MOSFET model, the noise PSD can be broken down into thermal and flicker noise components:

$$S_i = \frac{g_m^2 K_F}{L \cdot W \cdot C_{ox}} \cdot \frac{1}{f} + 4 \cdot k_B \cdot T \cdot \left(g_{mb} + \frac{2}{3} g_m \right), \tag{4}$$

where K_F is the flicker noise coefficient, C_{ox} is the gate oxide capacitance, f is the frequency, W and L are the transistor dimensions, k_B is the Boltzmann constant, T is the absolute temperature, g_{mb} accounts for substrate bias effects.

Since flicker noise ($1/f$ noise) dominates in low-frequency applications, its reduction is critical. Equation (4) reveals that flicker noise is inversely proportional to the transistor area ($W \times L$). Therefore, one of the key strategies that can be employed is increasing transistor dimensions. Using larger MOSFETs in the reference current source reduces flicker noise, but requires careful layout optimization to avoid excessive parasitic capacitance. By implementing this method, a lower noise floor can be achieved, enabling higher output impedance without compromising.

Analysis of noise compensation in feedback circuit

The differential amplifier is used to stabilize the voltage at the drain of the current sourcing transistor M44 in Fig. 1. The common mode voltage at the drain of transistor M44 or at the input of the Op-Amp is lower than the threshold voltage. Therefore, a folded cascode circuit of the Op-Amp is used. Let us consider the feedback analysis to show the effect of the Op-Amp noise on the output noise level of the proposed circuit. For this purpose, let us make a structural diagram of the stabilization circuit and its graph as shown in Fig. 2, *b* and *c*, respectively.

The expression for the transfer function of the Op-Amp noise and for the noise of the reference current source are obtained using Mason's formula.

$$T_{n_{OA}}(p) = \frac{T_{GS}(p)}{1 + T_{OA}(p) \cdot T_{GS}(p)} = \frac{1}{T_{OA}(p) + \frac{1}{T_{GS}(p)}}; \tag{5}$$

$$T_{n_{ref}}(p) = \frac{T_{OA}(p) \cdot T_{GS}(p)}{1 + T_{OA}(p) \cdot T_{GS}(p)} = \frac{1}{1 + \frac{1}{T_{OA}(p) \cdot T_{GS}(p)}}, \tag{6}$$

where $T_{GS}(p)$, $T_{OA}(p)$ are the transfer functions respects to the directed graph in the Fig. 2. In precision analog circuits, single-pole transfer functions are often employed to model system behavior. The DC transfer function of the transconductance stage ($T_{GS}(p)$) approximates unity, while the Op-Amp exhibits a DC gain typically exceeding 100. As a result, equation (5) demonstrates that the intrinsic noise of the Op-Amp is effectively suppressed within the closed-loop bandwidth of the control system. Consequently, its impact on the output noise current becomes negligible under normal operating conditions.

However, equation (6) reveals a critical limitation: the reference current source introduces noise components that appear unattenuated in the output current path. Consequently, this noise contribution becomes the primary limiting factor for the circuit's total noise characteristics.

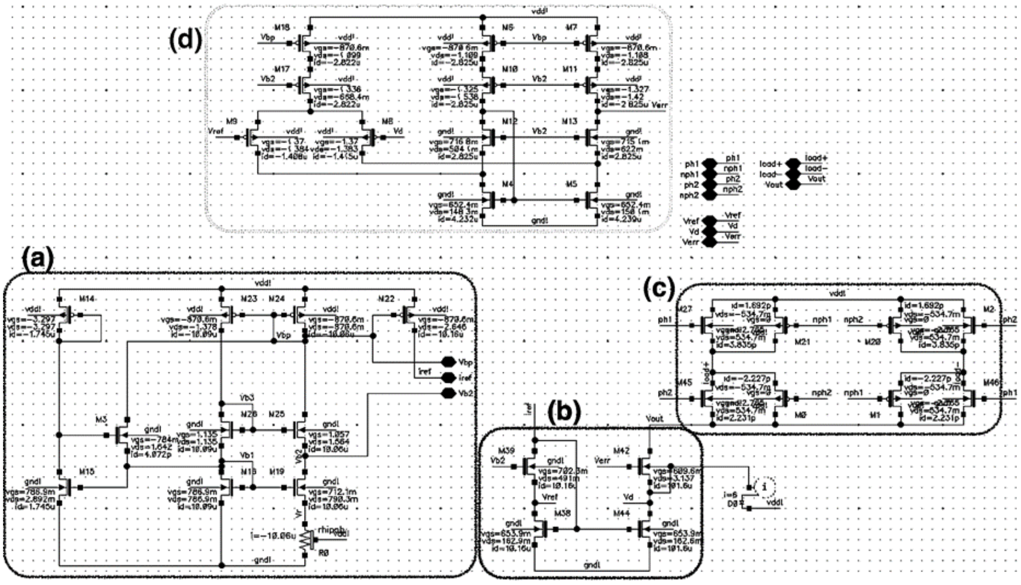


Fig. 1. Proposed reference current source circuit: start-up circuit (*a*), current driver (*b*), bipolar pulse forming circuit (*c*), Op-Amp (*d*)

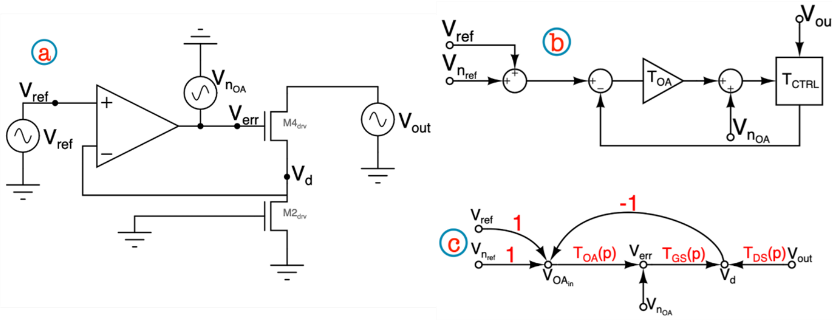


Fig. 2. Stabilizer circuit schematic (*a*), diagram (*b*) and graph (*c*)

In an ideal (noiseless) scenario, the voltage at node V_d (Fig. 2) is stabilized with an accuracy inversely proportional to the Op-Amp's gain. Thus, increasing the gain enhances output current stability, but only up to a certain point. Beyond this, circuit noise – primarily from the reference current source and thermal effects imposes a fundamental limit on stabilization precision.

Additionally, the frequency response of the Op-Amp plays a crucial role. Excessive bandwidth increases high-frequency noise, degrading signal integrity. Insufficient bandwidth restricts the circuit's ability to track and regulate current at the desired frequencies. Therefore, optimal Op-Amp design requires a careful balance: moderate gain selection (high enough to ensure stability, but not so high that noise dominates) and controlled bandwidth (wide enough to meet frequency requirements, but narrow enough to minimize noise amplification). By carefully considering these trade-offs, highly stable current outputs while minimizing noise-induced errors can be achieved.

Proposed estimation algorithm

The output current variations caused by the load impedance variations depend on the driver output impedance and are determined by the current divider formula (1). In turn, the noise-induced variations can be defined as the RMS noise current at the output of the circuit. Equating these two quantities gives the expression:

$$\Delta I_{R_L} = I_{n_{rms}} = \frac{\Delta R_L}{R_L + R_{out_{\max}}} \cdot I_{out}. \tag{7}$$

Expressing the output impedance from expression (7), we obtain the value of the feasible output impedance. As a good approximation, we can assume that the value of the load impedance varies from its typical value R_L to 0. In other words, assuming $\Delta R_L = R_L$ and that the output current is much larger than the RMS noise current, we obtain the following expression:

$$R_{out_{\max}} = \frac{I_{out}}{I_{n_{rms}}} \cdot \Delta R_L - R_L \approx \frac{I_{out}}{I_{n_{rms}}} \cdot R_L. \tag{8}$$

This formula can be used to estimate the required current driver output impedance for the given noise level at the driver output.

Design considerations using proposed estimation algorithm

The current driver designed in 180 nm process, using a 3.3 V supply voltage to ensure sufficient voltage headroom for driving high-impedance loads. This voltage selection was critical to maximize the achievable load resistance range and provide sufficient overdrive voltage for proper transistor operation.

As the reference current source constitutes the primary noise contributor in the system, our design methodology prioritized its optimization through: flicker noise reduction (which includes implementing large-area transistors, using pMOS devices for the reference branch, due to their superior flicker noise characteristics, and applying layout techniques, such as common-centroid placement, to reduce process variations) and architecture selection (adopted an Op-Amp-free architecture of the reference current source to reduce power consumption, minimize additional noise sources simplify circuit in micron process).

Therefore, a modified beta-multiplier reference with additional current mirroring transistors (M25, M26) have been implemented to improve output impedance, enhance power supply rejection and set operating points of the amplifier.

Simulation of a Current Driver in 180 nm CMOS Technology

Fig. 1, *a*, shows the schematic of the proposed reference current source. A parametric analysis was considered to make a trade-off between noise level and IC area. A transistor channel length of 3 µm was used. The transistor widths were chosen to give a nominal current of 10 µA.

This results in a noise RMS current of 3.4 nA. At start-up, the reference current source has two possible states: normal operation and zero current operation, like any self-biased circuit. To ensure normal operation, the start-up circuit shown in Fig. 1, *a* was added, consisting of transistors M14, M15 and M3. Fig. 3 shows the operation of the circuit on power-up without the start-up circuit (top diagram) and with the start-up circuit (bottom diagram).

The RMS value of the noise current can be used to evaluate the required output impedance using equation (8), and consequently the required gain of the differential amplifier using equation (2). The spectral density of the noise current at the output of the reference source is 3.4 nA for the circuit shown in Fig. 1, *a*. Considering that the current mirror has a gain factor $K = 10$, the RMS noise current at the output is more than 34 nA. The required output impedance according to (7) is roughly 30 MΩ. Taking into account the small signal parameters of the MOS transistor typical for 180 nm CMOS technology, we obtain the required gain of 56 dB. A folded cascode Op-Amp is used because the common mode voltage (V_{CM}) at the input of the Op-Amp is about 163 mV, well below the threshold voltage (V_{th}).

The DC analysis of the driver circuit reveals several interesting characteristics that validate the design approach. Fig. 4 shows the key performance metrics, presenting the output impedance

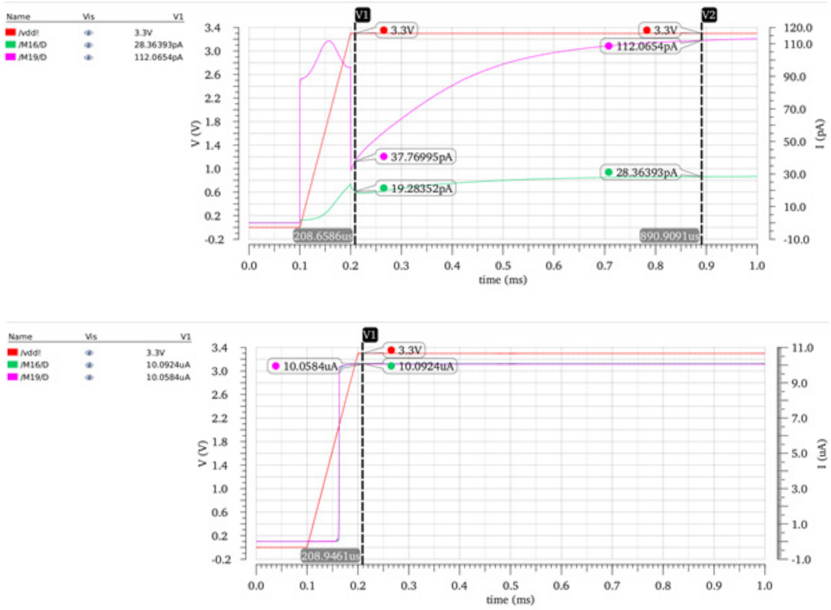


Fig. 3. Operation of the reference current source at power-up
without the start-up circuit (top) and with the start-up circuit (bottom)

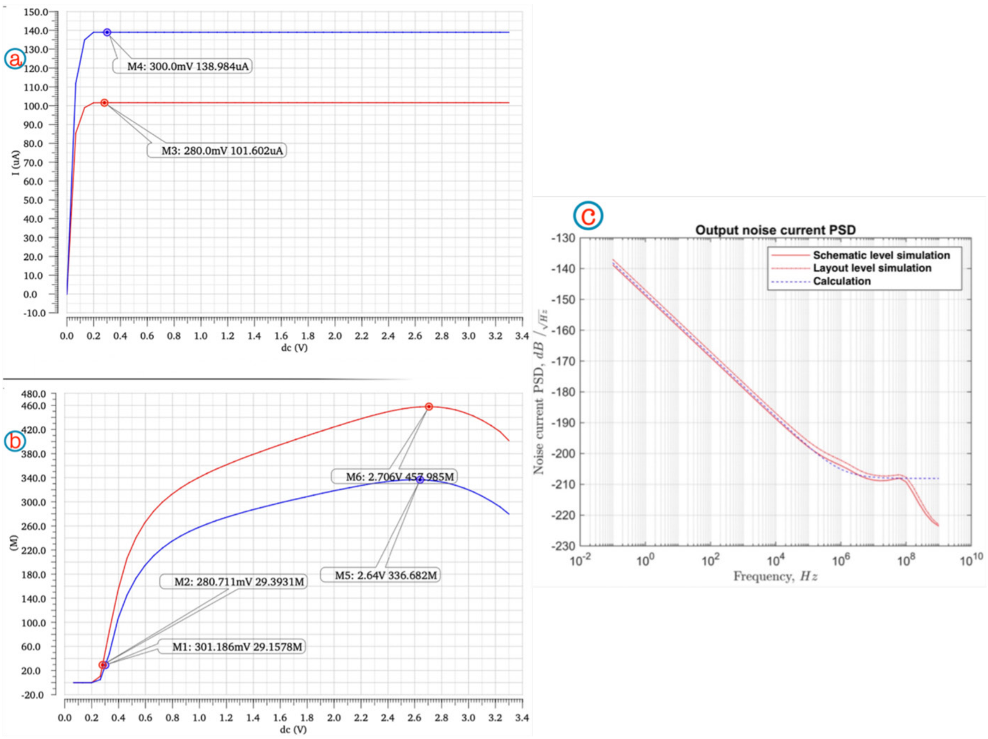


Fig. 4. Results of DC and noise analysis: I-V characteristics of the proposed current driver (*a*),
its output impedance (*b*), and deviation between calculated and simulated noise PSD (*c*).
Red curves represent circuit-level simulation, while blue curves correspond to layout-level simulation

characteristics in panel (Fig. 4, *a*) and comparing the calculated and simulated output noise current spectral densities in panel (Fig. 4, *b*). Note, there is close agreement between the schematic-level simulations (red curve) and the post-layout results (blue curve), with discrepancies not exceeding 2 dB

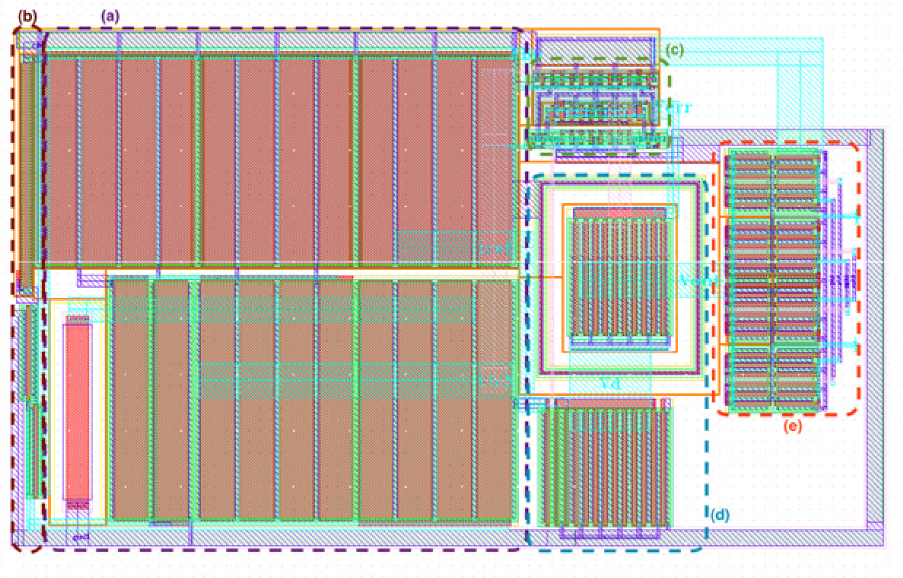


Fig. 5. Device layout: reference current source (*a*), start-up circuit (*b*), differential amplifier (*c*), current mirror ($K = 10$) (*d*), H-bridge (*e*)

within the frequency band of interest. This close correlation confirms that designers can reliably optimise the reference current source using these calculation methods, eliminating the need to simulate the complete circuit each time and significantly speeding up the design process.

The physical implementation is shown in Fig. 5 and demonstrates the robust performance of the circuit through several key specifications. The design generates a 140 μA output current from a 10 μA reference current while maintaining 90.9% voltage swing capability ranging from 300 mV to the full 3.3 V supply rail. The Op-Amp provides 54 dB of gain with a 190 kHz bandwidth, striking a balance between precision and speed. However, it should be noted that the output impedance exceeds requirements across much of the operational output voltage range. This performance margin stems from the inherent properties of the MOS transistors, as output impedance shows a strong dependence on overdrive voltage.

A more detailed examination of the circuit's behaviour reveals temperature stability, with output current variation of less than 0.1% per degree Celsius across typical temperature ranges (–40°C to 80°C). The amplifier maintains its 54 dB gain within 3 dB even at temperature extremes, demonstrating robust thermal performance.

Fig. 5 shows that transistors M42 and M44 are atypically arranged for cascading. This is due to the substrate current induced the body bias effect (SCBE) in the technology employed. To prevent the current from flowing into the substrate of transistor M42, its source and bulk terminals are shorted, and the transistor itself is isolated from the rest of the chip by a pn junction. This is the reason for the difference in the output current amplitude between schematic and layout level simulation.

It can be seen from the simulation results that the actual RMS noise at the output is more significant than ten times the noise of the reference current source assumed in the calculation of the rational output impedance, since only the contribution of the reference source has been taken into account. This provides a margin of output impedance in terms of feasibility with respect to noise level.

Comparison with the state of the art

Table 1 provides a detailed comparison of the key performance parameters of various implementations. References [1] and [4] have the highest output impedance, exceeding 1 GΩ, but require a large silicon footprint of over 0.2 mm^2. The ±9 V bipolar architecture in reference [5] enables superior current programmability and load handling capability, making it particularly versatile for stimulation applications. However, this comes at a cost: operational power dissipation reaches 70 mW, and the implementation occupies a similar area to that of reference [4]. Furthermore, the design is unable to sustain consistent output impedance across its full current range, which limits its precision in certain operating regimes.

Table 1

Comparison with the state of the art

	[1]	**[2], [6]**	**[4]**	**[5]**	**This paper**
CMOS process	180 nm	IBM 130 nm	IBM 130 nm	180 nm	180 nm
Supply	bipolar, 5 V	3.3 V	bipolar, 3.3 V	bipolar, 9 V	3.3 V
Output impedance (R_{out})	1 GΩ	100 kΩ	320 GΩ	–	30 MΩ
Voltage range at the output (as a percentage of the supply voltage)	90%	60%	90.9%	91.1% Vdd	90.9% Vdd
Current amplitude	from 0.02 mA to 5.1 mA	from 0.01 mA to 1 mA	0.096 mA	from 0.032 mA to 10 mA	0.14 mA
IC area	0.5 mm^2	0.015 mm^2	0.2 mm^2	0.19 mm^2	>0.035 mm^2
Power consumption stand-by / stimulation	–	– / 1 mW	– / 0.6 mW	1 μW / 70 mW	0.1 mW / 0.4 mW

Reference [2] is the area-optimized extreme at just 0.015 mm^2, though this minimization compromises other critical specifications, including output impedance and voltage compliance. The proposed design strikes a balance between competing parameters. It reduces power consumption by 24% compared to [4], while delivering a higher output current. With a die area of less than 0.035 mm^2, it occupies 3 times more space than reference [2], but remains 5 times more compact than other implementations. The solution maintains more than 90% voltage utilization, which is comparable to that of [4], while providing 30 MΩ output impedance, which is sufficient for target applications. Further increases would not improve current stability, but would negatively impact area, power and complexity. This analysis shows that the implemented architecture successfully balances the fundamental design trade-offs between performance, size and efficiency in current source implementations.

Conclusion

The developed noise analysis methodology enables the calculation of the required current driver parameters and establishes a practical upper limit of 30 MΩ for output impedance under typical operating conditions. This threshold was determined using a reference current source that generated 3.4 nA RMS noise at a nominal current of 10 μA. The analysis shows that increasing the impedance value towards 100 GΩ provides diminishing returns in terms of current stability, while substantially increasing silicon area and power requirements.

The implemented circuit demonstrates measurable improvements over existing solutions in three critical areas: power efficiency, die area utilization and operational voltage range. Experimental results confirm stable operation with 30 MΩ output impedance, maintaining over 90% supply voltage utilization (300 mV to 3.3 V) when delivering a 140 μA output current. These specifications translate to a maximum supported load resistance of approximately 21 kΩ at full output current. The complete implementation occupies less than 3500 μm^2 of silicon area.

This balanced approach prioritizes practical performance metrics over theoretical maximums, recognizing that the excessive pursuit of ultra-high output impedance provides negligible benefits for actual current stability, while incurring significant implementation complexity costs. Instead, the design methodology focuses on achieving sufficient noise-limited performance within constrained area and power budgets, making it particularly suitable for applications where these practical considerations outweigh purely theoretical performance benchmarks.

The voltage-dependent output impedance characteristics suggest opportunities for adaptive biasing approaches in future iterations, particularly for applications requiring wide output ranges. While originally designed for precision current delivery, the architecture shows potential for scaling to higher current applications or adaptation to specialized fields like biomedical instrumentation, where its combination of precision and stability would be particularly valuable.

REFERENCES

1. **Rozgić D., Hokhikyan V., Jiang W., Akita I., Basir-Kazeruni S., Chandrakumar H.** A 0.338 cm^3, artifact-free, 64-contact neuromodulation platform for simultaneous stimulation and sensing. *IEEE Transactions on Biomedical Circuits and Systems*, 2019, Vol. 13, No. 1, Pp. 38–55. DOI: 10.1109/TBCAS.2018.2889040

2. **Abdelhalim K., Jafari H.M., Kokarovtseva L., Perez Velazquez J.L., Genov R.** 64-channel UWB wireless neural vector analyzer SOC With a closed-loop phase synchrony-triggered neurostimulator. *IEEE Journal of Solid-State Circuits*, 2013, Vol. 48, No. 10, Pp. 2494–2510. DOI: 10.1109/JSSC.2013.2272952

3. **Lee J., Rhew H.-G., Kipke D.R., Flynn M.P.** A 64 channel programmable closed-loop neurostimulator with 8 channel neural amplifier and logarithmic ADC. *IEEE Journal of Solid-State Circuits*, 2010, Vol. 45, No. 9, Pp. 1935–1945. DOI: 10.1109/JSSC.2010.2052403

4. **Maghami M.H., Sodagar A.M., Sawan M.** Analysis and design of a high-compliance ultra-high output impedance current mirror employing positive shunt feedback. *International Journal of Circuit Theory and Applications*, 2015, Vol. 43, No. 12, Pp. 1935–1952. DOI: 10.1002/cta.2049

5. **Haas M., Vogelmann P., Ortmanns M.** A neuromodulator frontend with reconfigurable Class-B current and voltage controlled stimulator. *IEEE Solid-State Circuits Letters*, 2018, Vol. 1, No. 3, Pp. 54–57. DOI: 10.1109/LSSC.2018.2827885

6. **Kassiri H., Bagheri A., Soltani N., Abdelhalim K., Jafari H.M., Salam M.T.** Battery-less tri-band-radio neuro-monitor and responsive neurostimulator for diagnostics and treatment of neurological disorders. *IEEE Journal of Solid-State Circuits*, 2016, Vol. 51, No. 5, Pp. 1274–1289. DOI: 10.1109/JSSC.2016.2528999

7. **Wang Y., Luo H., Chen Y., Jiao Z., Sun Q., Dong L.** A closed-loop neuromodulation chipset with 2-level classification achieving 1.5-Vpp CM interference tolerance, 35-dB stimulation artifact rejection in 0.5ms and 97.8%-sensitivity seizure detection. *IEEE Transactions on Biomedical Circuits and Systems*, 2021, Vol. 15, No. 4, Pp. 802–819. DOI: 10.1109/TBCAS.2021.3102261

8. **Wang A., Jung D., Park J., Junek G., Wang H.** Electrode–electrolyte interface impedance characterization of ultra-miniaturized microelectrode arrays over materials and geometries for sub-cellular and cellular sensing and stimulation. *IEEE Transactions on NanoBioscience*, 2019, Vol. 18, No. 2, Pp. 248–252. DOI: 10.1109/TNB.2019.2905509

9. **Lempka S.F., Miocinovic S., Johnson M.D., Vitek J.L., McIntyre C.C.** In vivo impedance spectroscopy of deep brain stimulation electrodes. *Journal of Neural Engineering*, 2009, Vol. 6, No. 4, Art. no. 046001. DOI: 10.1088/1741-2560/6/4/046001

10. **Maghami M.H., Sodagar A.M., Sawan M.** Versatile stimulation back-end with programmable exponential current pulse shapes for a retinal visual prosthesis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2016, Vol. 24, No. 11, Pp. 1243–1253. DOI: 10.1109/TNSRE.2016.2542112

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Mironov Kirill A.
Миронов Кирилл Александрович
E-mail: mironov.ka@edu.spbstu.ru
ORCID: https://orcid.org/0009-0001-8226-6288

Morozov Dmitry V.
Морозов Дмитрий Валерьевич
E-mail: dvmorozov@inbox.ru
ORCID: https://orcid.org/0000-0003-3403-0120

Akhmetov Denis B.
Ахметов Денис Булатович
E-mail: akhmetov@spbstu.ru
ORCID: https://orcid.org/0000-0002-1291-0584

Submitted: 08.11.2024; Approved: 14.02.2025; Accepted: 19.05.2025.

Поступила: 08.11.2024; Одобрена: 14.02.2025; Принята: 19.05.2025.

Short message
DOI: https://doi.org/10.18721/JCSTCS.18209
UDC 621.3.049.774.2

A PIPELINE ANALOG-TO-DIGITAL CONVERTER IN 180 NM CMOS

M.M. Pilipko ✉ , *D.V. Morozov*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ m_m_pilipko@rambler.ru

Abstract. A pipelined analog-to-digital converter (ADC) is presented, which was designed using 180 nm complementary metal-oxide semiconductor (CMOS) technology with a supply voltage of 1.8 V from Micron JSC. The ADC circuit consists of a sample-and-hold device, an 8-level redundant stage, five 6-level redundant pipeline stages, a back-end 3-bit ADC, as well as synchronization circuits, an adder and multiplexers to get at the output the 16-bit direct binary code of the whole ADC or the redundant code from first to fifth stages. The pipeline is implemented as a switched-capacitor circuit using operational transconductance amplifiers. The simulation of the ADC in the time domain in the Virtuoso analog design environment from Cadence DS was performed. The clock frequency was set to 50 MHz. The power consumption was about 52 mW, the following main characteristics were achieved: SINAD = 74.6 dB (ENOB = 12 bits) and SFDR = 75.3 dB.

Keywords: analog-to-digital converter, pipeline ADC, bootstrapped switch, time interleaving, redundant stage

Acknowledgements: The production of the integrated microcircuit was funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of the federal project “Training of personnel and scientific foundation for the electronic industry” under the state assignment for the implementation of the research work “Development of a methodology for prototyping an electronic component base in domestic microelectronic production based on the MPW service”.

Citation: Pilipko M.M., Morozov D.V. A pipeline analog-to-digital converter in 180 nm CMOS. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 111–119. DOI: 10.18721/JCSTCS.18209

Краткое сообщение
DOI: https://doi.org/10.18721/JCSTCS.18209
УДК 621.3.049.774.2

КОНВЕЙЕРНЫЙ АНАЛОГО-ЦИФРОВОЙ ПРЕОБРАЗОВАТЕЛЬ ПО ТЕХНОЛОГИИ КМОП 180 НМ

М.М. Пилипко ✉, *Д.В. Морозов*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ m_m_pilipko@rambler.ru

Аннотация. Представлен конвейерный аналого-цифровой преобразователь (АЦП), который выполнен по 180 нм комплементарной металл-оксид-полупроводник (КМОП) технологии с напряжением питания 1,8 В от компании АО «Микрон». Схема АЦП состоит из устройства выборки и хранения, каскадов с избыточностью (8 уровней квантования в первом каскаде, 6 уровней в каскадах 2–6), оконечного АЦП с разрядностью 3 бита, а также схем синхронизации, сумматора и мультиплексоров для вывода либо прямого 16-разрядного двоичного кода, либо кода каскадов с избытком. Конвейер реализован как схема на переключаемых конденсаторах с использованием операционных транскондуктивных усилителей. Моделирование АЦП во временной области проводилось в среде аналогового проектирования Virtuoso компании Cadence DS. Тактовая частота была равна 50 МГц. Потребляемая мощность составила около 52 мВт, были достигнуты следующие основные характеристики: SINAD = 74,6 дБ (ENOB = 12 бит) и SFDR = 75,3 дБ.

Ключевые слова: аналого-цифровой преобразователь, конвейерный АЦП, ключ с постоянным сопротивлением, временное перемежение, каскад с избыточностью

Финансирование: Производство интегральной микросхемы было выполнено за счет средств Министерства науки и высшего образования Российской Федерации в рамках федерального проекта «Подготовка кадров и научного фундамента для электронной промышленности» по государственному заданию на выполнение научно-исследовательской работы «Разработка методики прототипирования электронной компонентной базы на отечественных микроэлектронных производствах на основе сервиса MPW».

Для цитирования: Pilipko M.M., Morozov D.V. A pipeline analog-to-digital converter in 180 nm CMOS // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 111–119. DOI: 10.18721/JCSTCS.18209

Introduction

For the implementation of precision high-speed analog-to-digital converters (ADCs), the most promising one is the currently widely used pipeline architecture. On one hand, pipelined ADCs provide a higher resolution with a lower power consumption compared to flash ADCs. On the other hand, pipelined ADCs operate at higher frequencies compared to successive-approximation ADCs [1]. Therefore, these ADCs offer a compromise between the achieved resolution and the operating frequency. The pipelined ADCs are used in applications, such as high-definition video, wireless local area networks etc. [2, 3].

A typical pipelined ADC consists of a sample-and-hold (S/H) device and a number of ADC stages connected in series one after another. The first stage of the pipelined ADC allows to determine the most significant bits of the ADC output code. The next stages of the pipeline sequentially determine the next significant bits of the ADC output code. A stage contains a low-resolution ADC (usually a flash ADC of 1–4 bits) and a multiplying digital-to-analog converter (DAC). One of the possible approaches to realize the DAC is the use of a switched-capacitor circuit.

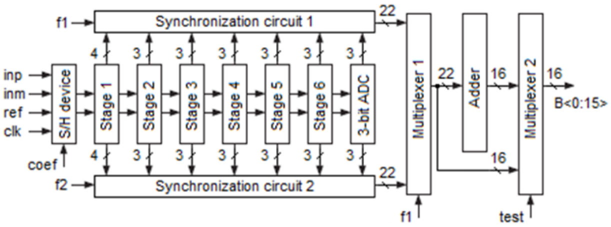


Fig. 1. Block diagram of the pipelined ADC

The paper presents a block diagram of the designed pipelined ADC and discusses the implementation of its blocks based on a complementary metal-oxide-semiconductor (CMOS) technology. A specialized computer-aided design system was used to obtain simulation results and define the main characteristics of the pipelined ADC.

Block diagram of the pipeline ADC

Fig. 1 shows a block diagram of the pipelined ADC. The differential input signal ("inp" and "inm") is fed to the S/H device, whose gain is discretely set as 1 or 2 by the signal "coef". The clock frequency of the S/H device is set by the clock signal (clk). The signal "ref" corresponds to the voltage of the direct current (DC) operating point for the input signals ("inp" and "inm"). The S/H device can process both a differential input signal and the non-differential signal at the input "inp", in the latter case the signal "ref" should be fed to the input "inm".

The differential signal from the output of the S/H device is fed to sequentially connected pipeline stages. In the pipeline, interleaving is organized within two half-periods of the clock signal, which are designated as phases "f1" and "f2". The first stage of the pipeline has a 4-bit output from the flash ADC with 8 comparators. Each of the following five stages has a 3-bit output from its flash ADC with 6 comparators. The output of the sixth stage is fed to the back-end flash ADC with 7 comparators and 3 output bits. Compared to a basic 2-bit stage, the implemented pipeline stages provide redundancy in the representation of the output code. The output signals of the stages and the 3-bit ADC are delayed by the required number of clock signal periods using synchronization circuits 1 and 2. As a result, codes that correspond to a respective input sample are properly arranged.

The signals from the outputs of the synchronization circuits are fed to the multiplexer 1, which is controlled by phase "f1" at the address input. When the phase signal "f1" takes the value 1, the output signal of the synchronization circuit 1 appears at the output of the multiplexer 1. During the time when the phase signal "f1" takes the value 0 (phase "f2" is actually in effect), the output signal of the synchronization circuit 2 appears at the output of the multiplexer 1. This way the interleaving is realized in the digital part of the circuit.

The signal from the output of multiplexer 1 is fed to the adder, which performs summation taking into account the weight of the redundant codes at its inputs and generates the ADC output code. The multiplexer 2 at the ADC output, which is controlled by the test signal (test) at the address input, allows either the straight binary code from the adder output or 16 digits of the redundant code from first to fifth stages to pass to the circuit output. The latter option allows for off-chip correction of the interstage gain error and other ADC imperfections.

Cadence Virtuoso circuit of the pipelined ADC

The pipelined ADC based on switched capacitors has been developed in the computer-aided design system Virtuoso by Cadence DS, where the 180 nm CMOS technology with a supply voltage of 1.8 V from Mikron JSC was used. The layout of the pipelined ADC is shown in Fig. 2. Sizes of the layout are 1600×300 um.

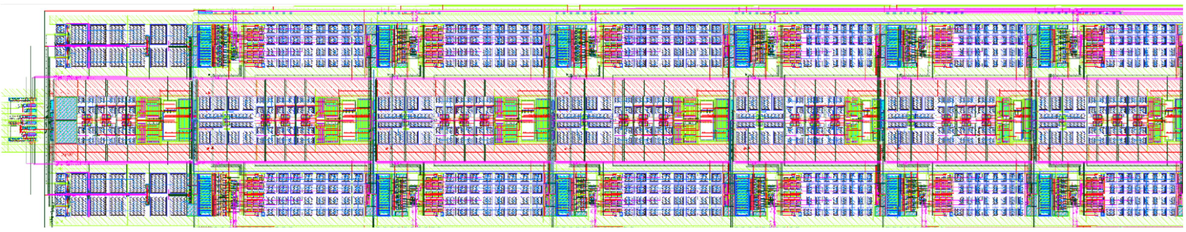


Fig. 2. Layout of the pipelined ADC

The circuit of the S/H device is shown in Fig. 3. Input signals (non-inverting "inp" and inverting "inm") are applied to the circuit via bootstrapped switches [4, 5]. The S/H device is based on a folded-cascode rail-to-rail operational transconductance amplifier (OTA) similar to [6, 7], but with additional boosting amplifiers to increase the gain according to recommendations [8, 9]. According to the simulation results at 27°C, at a 7 pF load the DC gain of OTA1 is 91.5 dB, the unity-gain bandwidth is 335 MHz, the phase margin is 57 degrees, power consumption is 9.2 mW. The OTA inputs are non-inverting (vp) and inverting (vm), the OTA outputs are non-inverting (vop) and inverting (vom). The supply voltage is designated as "vdd!". The circuit uses CMOS switches controlled by two non-overlapping phases "f1" and "f2" similar to [6]. Phases "nf1" and "nf2" are inverted to "f1" and "f2", respectively. The S/H device outputs "outp" and "outm" are taken from the OTA outputs "vom" and "vop", respectively.

Capacitors with a nominal value of 1.8 pF are connected in the feedback loops of OTA1. When the signal "coef" takes the value 0, the input capacitance of the S/H device consists of capacitors with a nominal value of 1.8 pF to both non-inverting (inp) and inverting (inm) inputs. This ratio of input and feedback capacitances sets the unity gain in the S/H device. When the signal "coef" takes the value 1, the input capacitance of the S/H device consists of two parallel-connected capacitors with a nominal value of 1.8 pF to both non-inverting (inp) and inverting (inm) inputs, which totally give 3.6 pF in each of the specified nodes. This sets a 2x gain in the S/H device. As seen, the OTA is used by the circuit only in phase "f1". In phase "f2", the OTA is used by an identical S/H device that samples the input in phase "f1". This principle of operation allows for interleaving and effectively doubles the conversion rate of the ADC.

As said before, the S/H device [10] can convert a non-differential signal to a differential signal, which is preferable for the further analog-to-digital conversion. As can be seen from Fig. 3, there is a CMOS switch between the nodes "sw0p" and "sw0m" and a CMOS switch between the nodes "sw1p" and "sw1m". Assume that a harmonic signal is fed to "inp", while the signal "ref" (half the supply voltage "vdd!") is fed to "inm". The signal "coef" is set to 1. In the sample phase, the capacitors C1 and C2 have the same charge Qinp, which is different from the charge of capacitors C3 and C4. The latter ones have the voltage "ref" at both plates, thus, each of the capacitors C3 and C4 has the same charge equal to 0. In the hold phase, CMOS switches between nodes "sw0p" and "sw0m", "sw1p" and "sw1m" are closed, and the Qinp charge of C1 is divided equally between capacitors C1 and C4, while the Qinp charge of C2 is divided equally between capacitors C2 and C3. Therefore, all capacitors have the same charge equal to Qinp/2. This coefficient of 1/2 is compensated by the 2x gain of the S/H device.

The S/H device was simulated at 27°C in the time domain. Simulation results for the case, when the S/H device converts a non-differential signal to a differential signal, are shown in Fig. 4. The clock frequency is set to 50 MHz. The input harmonic signal frequency at the input (inp) is set to 781.25 kHz, the amplitude is 600 mV. The signal "ref" is 900 mV DC. The S/H device output signals "vom" and "vop" are depicted in the second plot.

The circuit of the first stage in the pipeline is shown in Fig. 5. The resistive divider R0-R14 between the power supply node "vdd!" and the ground node "gnd!" sets 8 reference voltages from 375 mV to

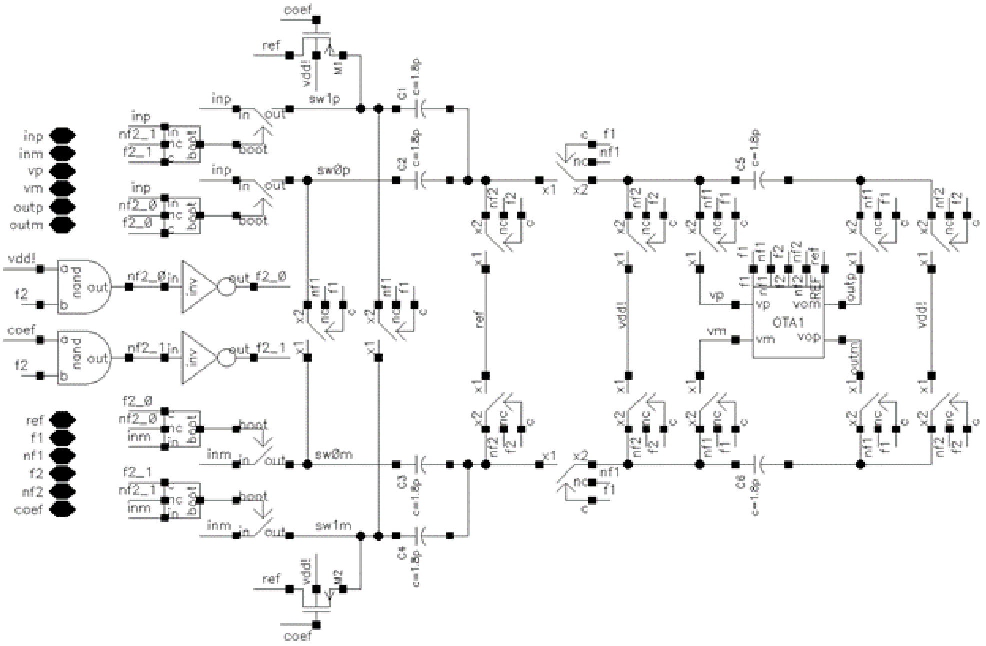


Fig. 3. S/H device in Cadence Virtuoso

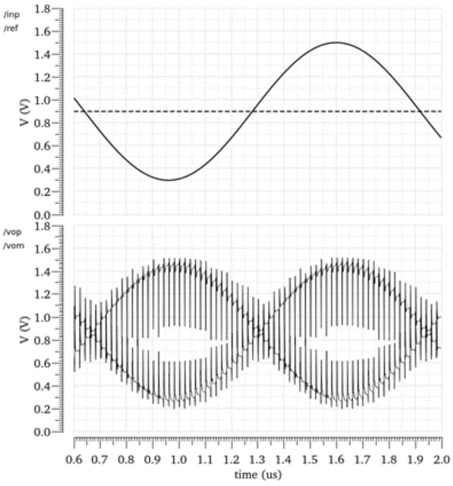


Fig. 4. Simulation results of the S/H device for the non-differential input signal

1425 mV in nodes “res<0:7>” for the comparators “comp” (Fig. 5, *a*). An excessive number of comparators allows for an extended voltage swing of the input signal that is limited by voltage levels 150 mV and 1.65 V. Differential comparator circuits similar to [6] are used. Input signals “inp” and “inm” are applied to the comparator inputs “vp” and “vm”, respectively. The reference voltages are fed to comparator inputs “vp2” and “vm2”. At the comparator outputs “out” and “nout”, direct and inverted codes of the comparison result are formed. The output thermometer code of the comparators is designated “t<0:7>” (the inverted code is “nt<0:7>”). The thermometer code is converted into a 4-bit direct binary output code of the first stage by an adder.

The input signals “inp” and “inm” are also applied to the switched-capacitor part of the first stage via bootstrapped switches (Fig. 5, *b*). The circuit also uses CMOS switches controlled by two phases

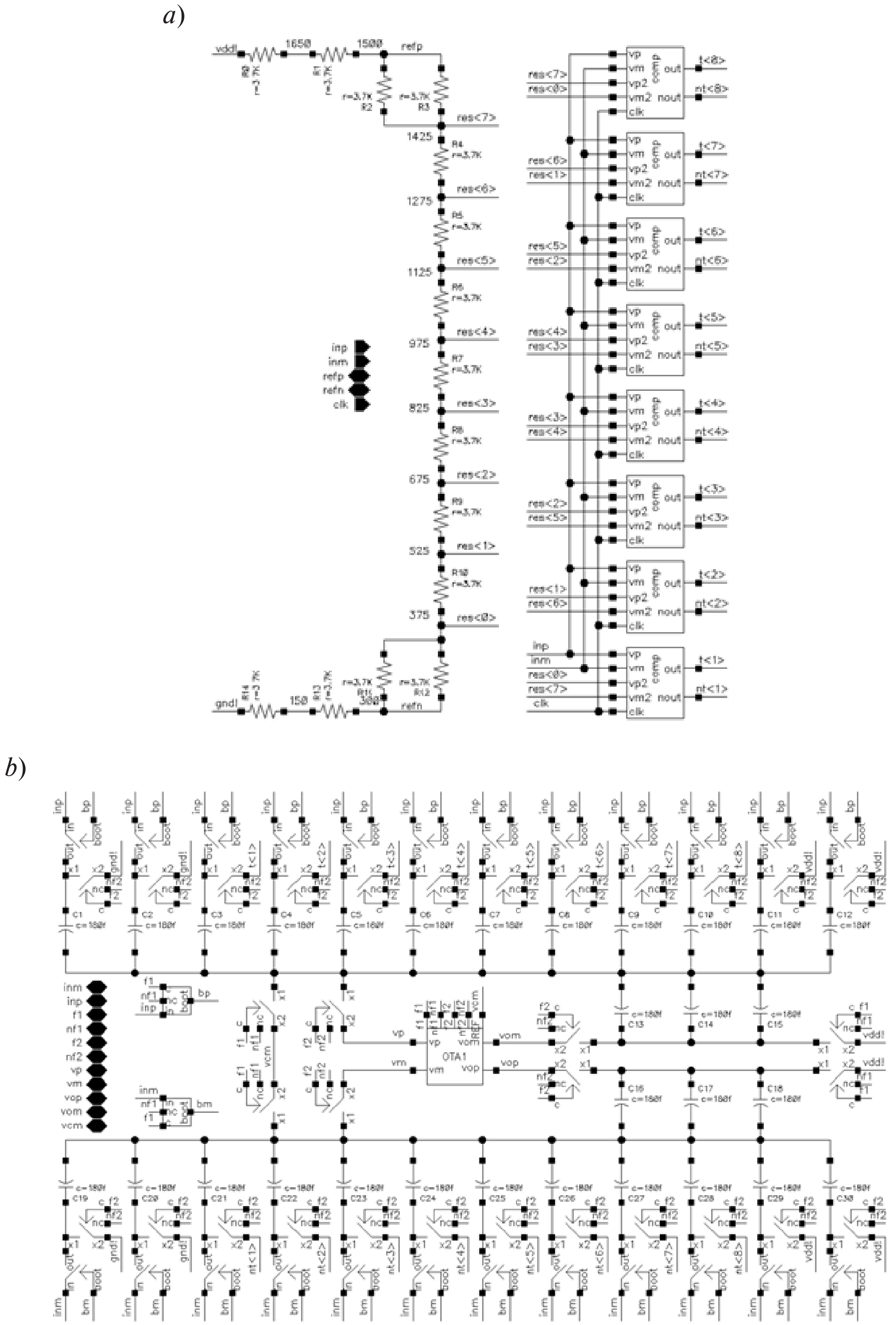


Fig. 5. First stage in Cadence Virtuoso

“f1” and “f2” (inverted “nf1” and “nf2” respectively). The first stage outputs are designated as “vop” and “vom”. The input capacitance of the first stage consists of 12 parallel-connected capacitors with a nominal value of 180 fF to both “inp” and “inm”, which totally give 2.16 pF in each of the specified nodes. In the feedback loops of OTA1 there are three capacitors with a nominal value of 180 fF connected in parallel, which totally give 540 fF. This ratio of capacitances sets a 4x gain in the first stage. Again, here the OTA is used by the circuit only in phase “f2”. In phase “f1”, this OTA is used by the interleaved ADC stage that samples the input in phase “f2”.

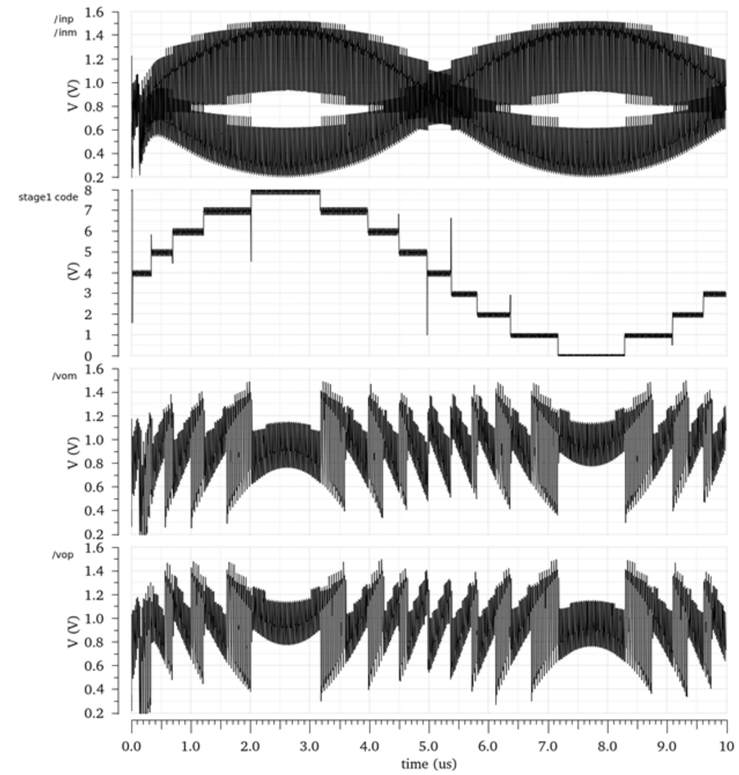


Fig. 6. Simulation results of the first stage

The output thermometer code of the comparators “t<0:7>” and “nt<0:7>” is fed to 8 of 12 CMOS switches in the corresponding branch of the switched capacitor circuit. The rest 4 of 12 switches are connected as follows: 2 switches – to the ground node “gnd!”, 2 switches – to the power supply node “vdd!”, which defines the voltage range of the DAC in the first stage between 300 mV and 1.5 V.

The first stage simulation results at 27°C in the time domain are depicted in Fig. 6. The differential input signal “inp” and “inm” of the first stage of the pipeline ADC is shown in the first plot. The second plot shows the 4-bit direct binary output of the first stage in decimal equivalent. At the output of the stage, a difference is formed between the differential input signal of the stage and the decimal equivalent of the output code. Two last plots show the output signals “vom” and “vop”, which are the input signals for the second stage of the pipelined ADC.

The remaining stages 2–6 of the pipelined ADC are similar to the first stage, the only differences are as follows. Each of the stages 2–6 has 6 comparators to form the 3-bit output code. The resistive divider R0-R14 (Fig. 5, *a*) sets 6 reference voltages from 525 mV to 1275 mV at nodes “res<1:6>” for the comparators. The output thermometer code of the comparators “t<1:6>” and “nt<1:6>” is fed to 6 of 12 CMOS switches in the corresponding branch of the switched-capacitor circuit (Fig. 5, *b*). The rest 6 of 12 switches are connected as follows: 3 switches – to the ground node “gnd!”, 3 switches – to the power supply node “vdd!”, which defines the operating voltage range of the DAC between 450 mV and 1350 mV. The stages 2–4 are based on OTA1, while stages 5 and 6 utilize OTA2, with the width of the transistors being two times less than the width of the transistors in OTA1. The back-end 3-bit ADC has a resistive divider that sets 7 reference voltages from 450 mV to 1350 mV, while 7 comparators form the 3-bit output code.

Power consumption of the pipelined ADC is near 52 mW. The output code spectrum of the pipelined ADC is shown in Fig. 7, a 512-point discrete Fourier transform with a rectangular window was performed. The calculated signal-to-noise and distortion ratio (SINAD) is 74.6 dB, i.e. the effective

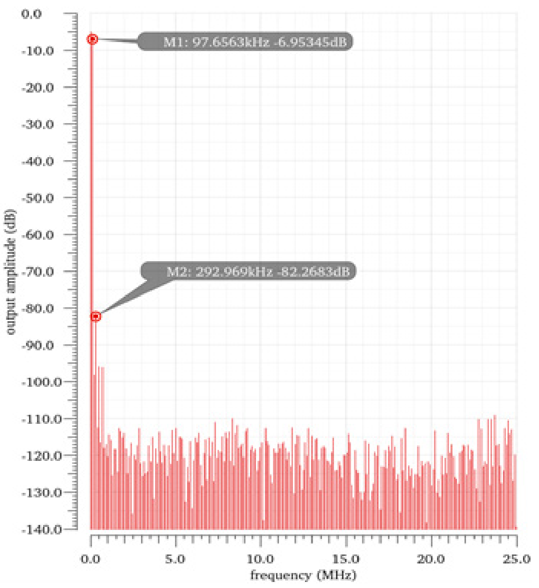


Fig. 7. Simulation results of the output code spectrum for the pipeline ADC

number of bits (ENOB) is 12 bits, spurious-free dynamic range (SFDR) is 75.3 dB. At a clock frequency of 5 MHz, SINAD is 78.8 dB, SFDR is 80.1 dB.

Conclusions

The pipelined ADCs are in demand in high-resolution high-speed applications. A 16-bit pipelined ADC was presented for 180 nm CMOS technology with a supply voltage of 1.8 V from Mikron JSC. Time interleaving was applied to effectively utilize the analog components and double the conversion rate. Pipeline stages use redundancy to compensate for the offset of comparators. Raw code of the pipeline stages is available at the outputs in the “test” mode, which makes further correction possible. The simulation of the ADC in time domain in the Virtuoso analog design environment from Cadence DS was performed. The clock frequency was set to 50 MHz. Power consumption was near 52 mW. For an input amplitude of 600 mV, SINAD = 74.6 dB (ENOB = 12 bits) and SFDR = 75.3 dB were achieved.

REFERENCES

1. **Pan H.** A/D converter fundamentals and trends. *2017 IEEE Custom Integrated Circuits Conference* (*CICC*), 2017, Pp. 1–102. DOI: 10.1109/CICC.2017.7993723

2. **Greeshma R., Anoop V.K., Venkataramani B.** A novel opamp and capacitor sharing 10 bit 20 MS/s low power pipelined ADC in 0.18µm CMOS technology. *2017 IEEE Computer Society Annual Symposium on VLSI* (*ISVLSI*), 2017, Pp. 594–599. DOI: 10.1109/ISVLSI.2017.110

3. **Hekimyan A., Bulakh D., Sahakyan A.** High accuracy pipelined ADC design for wireless LANs. *2015 Internet Technologies and Applications* (*ITA*), 2015, Pp. 312–314. DOI: 10.1109/ITechA.2015.7317415

4. **Razavi B.** The bootstrapped switch [A circuit for all seasons]. *IEEE Solid-State Circuits Magazine*, 2015, Vol. 7, No. 3, Pp. 12–15. DOI: 10.1109/MSSC.2015.2449714

5. **Kim S.-H., Lee Y.-H., Chung H.-J., Jang Y.-C.** A bootstrapped analog switch with constant on-resistance. *IEICE Transactions on Electronics*, 2011, Vol. E94.C, No. 6, Pp. 1069–1071. DOI: 10.1587/transele.E94.C.1069

6. **Pilipko M.M., Morozov D.V., Yenuchenko M.S.** MASH 2-2 delta-sigma modulator with dynamic element matching in 0.18 µm CMOS technology. *Computing, Telecommunications and Control*, 2023, Vol. 16, No. 3, Pp. 29–38. DOI: 10.18721/JCSTCS.16303

7. **Pilipko M.M., Morozov D.V., Yenuchenko M.S.** Delta-sigma modulator with 10 MHz clock frequency in 180 nm CMOS technology. *MES-2018*, 2018, Pp. 44–48. DOI: 10.31114/2078-7707-2018-4-44-48

8. **Baker R.J.** *CMOS: Circuit Design, Layout, and Simulation* (*IEEE Press Series on Microelectronic Systems*), 4th ed. Hoboken, New Jersey: Wiley-IEEE Press, 2019. 1235 p.

9. **Piatak I.M., Morozov D.V., Pilipko M.M.** A 14-bit 100 MS/s Pipelined ADC. *MES-2016*, 2016, Pp. 13–16.

10. **Zadeh A.** A 100MHz, 1.2V, ±1V peak-to-peak output, double-bus single ended-to-differential switched-capacitor amplifier for multi-column CMOS image sensors. *2016 14th IEEE International New Circuits and Systems Conference* (*NEWCAS*), 2016, Pp. 1–4. DOI: 10.1109/NEWCAS.2016.7604739

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Pilipko Mikhail M.
Пилипко Михаил Михайлович
E-mail: m_m_pilipko@rambler.ru
ORCID: https://orcid.org/0000-0003-3813-6846

Morozov Dmitry V.
Морозов Дмитрий Валерьевич
E-mail: dvmorozov@inbox.ru
ORCID: https://orcid.org/0000-0003-3403-0120

Submitted: 21.03.2025; Approved: 22.05.2025; Accepted: 30.05.2025.
Поступила: 21.03.2025; Одобрена: 22.05.2025; Принята: 30.05.2025.

Research article
DOI: https://doi.org/10.18721/JCSTCS.18210
UDC 004.896

A SOFTWARE SYSTEM FOR SURROGATE-BASED PROTOTYPING OF GAS TURBINE BLADES USING SERVERLESS CONTAINERS IN THE CLOUD

G.A. Zhemelev ✉ , *P.D. Drobintsev*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ wws.dev@gmail.com

Abstract. Design optimization of gas turbine blades is a complex multidisciplinary task requiring computationally expensive physics simulations. To perform them, a multitude of computer-aided engineering tools are used, often with machine-learning surrogates for rapid prototyping, all integrated into the optimization cycle. However, current approaches to such integration are hindered by the need for labor-intensive manual setups, vendor lock-in and a lack of scalable, automated workflows. We present a novel cloud-based architecture for building flexible optimization pipelines using containerized components. The proposed solution employs serverless containers, asynchronous messaging and cloud services to ensure the system's scalability, portability and resilience. Additionally, it follows MLOps principles to achieve reproducibility and efficient lifecycle management of machine learning models used in the optimization process. Unlike existing frameworks, our solution minimizes user setup complexity, allows easy integration of various software into the optimization cycle, and avoids vendor lock-in through open-source technologies and standard cloud APIs. Experiments with aerodynamic design optimization of gas turbine blades demonstrate the system's scalability, fault tolerance and successful integration of surrogate models for rapid blades prototyping. Furthermore, the system's flexibility and extensible architecture make it applicable to a broader range of engineering design optimization tasks beyond gas turbine blade aerodynamics.

Keywords: gas turbine blades, engineering design optimization, serverless containers, cloud computing, surrogate models, machine learning, MLOps

Citation: Zhemelev G.A., Drobintsev P.D. A software system for surrogate-based prototyping of gas turbine blades using serverless containers in the cloud. Computing, Telecommunications and Control, 2025, Vol. 18, No. 2, Pp. 120–136. DOI: 10.18721/JCSTCS.18210

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.18210
УДК 004.896

ПРОГРАММНАЯ СИСТЕМА ДЛЯ ПРОТОТИПИРОВАНИЯ ЛОПАТОК ГАЗОВЫХ ТУРБИН С ИСПОЛЬЗОВАНИЕМ СУРРОГАТНЫХ МОДЕЛЕЙ И БЕССЕРВЕРНЫХ КОНТЕЙНЕРОВ В ОБЛАКЕ

Г.А. Жемелев ✉, *П.Д. Дробинцев*

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ wws.dev@gmail.com

Аннотация. Оптимизация конструкции лопаток газовых турбин – это сложная мультидисциплинарная задача, требующая ресурсоемких физических расчетов. Для их выполнения применяют множество инженерных программных пакетов, часто вместе с суррогатными моделями машинного обучения в целях быстрого прототипирования. Однако на данный момент эффективная интеграция широкого спектра программного обеспечения (ПО) в цикле оптимизации остается актуальной проблемой ввиду трудоемкости установки и настройки компонентов, привязки к конкретным поставщикам ПО, а также недостатка масштабируемости и автоматизации вычислительных конвейеров. В данной статье предлагается оригинальная архитектура системы, основанная на использовании облачных сервисов и контейнеризованных компонентов с целью построения гибких вычислительных конвейеров для инженерной оптимизации. Предлагаемое решение включает в себя применение бессерверных вычислений на контейнерах и асинхронный обмен сообщениями, что вместе с использованием типовых облачных ресурсов позволяет обеспечить масштабируемость, переносимость и устойчивость системы. Кроме того, в ней применяется подход MLOps для эффективной организации жизненного цикла суррогатных моделей, что повышает качество и повторяемость результатов машинного обучения. Предложенное решение превосходит существующие благодаря простоте интеграции разнообразного ПО в цикле оптимизации и простоте в установке для пользователей, а также минимизирует зависимость от конкретных поставщиков за счет использования только открытого и свободно распространяемого ПО и стандартных облачных ресурсов. Проведенные эксперименты по оптимизации аэродинамики лопаток газовых турбин позволили убедиться в масштабируемости и отказоустойчивости системы, а также в ее применимости для быстрого прототипирования с использованием суррогатных моделей. Более того, гибкость разработанной системы и расширяемость ее архитектуры открывают возможности по применению предложенного решения и в других задачах инженерной оптимизации, не ограничиваясь проектированием лопаток газовых турбин.

Ключевые слова: лопатки газовых турбин, инженерное проектирование и оптимизация, бессерверные контейнеры, облачные вычисления, суррогатные модели, машинное обучение, MLOps

Для цитирования: Zhemelev G.A., Drobintsev P.D. A software system for surrogate-based prototyping of gas turbine blades using serverless containers in the cloud // Computing, Telecommunications and Control. 2025. Т. 18, № 2. С. 120–136. DOI: 10.18721/JCSTCS.18210

Introduction

Blades are key components of a gas turbine: their shape heavily affects the efficiency of extracting useful work from the gas flow and ultimately the performance of the entire energy generation unit. Finding the optimal shape of blades for each turbine stage is a time-consuming procedure that requires complex and resource-intensive calculations that model the physical processes in and around the blades.

The multitude of disciplines involved and the need to optimize hundreds of parameters that define the geometry and other properties of each blade – all add up to the labor and time required to bring new turbine models to market. Therefore, the industry is looking for ways to speed up computations, in particular, by using machine learning (ML) to construct surrogate models [1], as well as to increase the degree of automation of the entire blade prototyping process by integrating a variety of software into the continuous optimization cycle. This, in turn, leads to the need to organize the interaction of computer systems that participate in the design optimization of gas turbine blades. Fig. 1 shows a simplified workflow of MultiDisciplinary Optimization (MDO) of a gas turbine blade. Arrows indicate data flows and "P" blocks stand for optional postprocessing routines for each stage.

We can divide the depicted systems into two groups: inside and outside the optimization loop. Steps inside (on a gray background in Fig. 1) form a pipeline, which may include parallel steps. Surrogate models usually replace the pipeline by approximating functions that map design variables into objectives and constraints.

Surrogate models, also known as meta-models or meta-functions, are mathematical models that approximate the behavior of a complex system or function in order to speed up computation or simplify analysis by replacing the original model in relevant problems [2, 3]. Typically, the construction of surrogate models is based on experimental data and is implemented using ML techniques[1] [2–4]. Computational Fluid Dynamics (CFD), Conjugate Heat Transfer (CHT) and Finite Element Analysis (FEA) are perfect examples of disciplines involving long and costly computations, for which surrogate modeling can be used in the process of optimizing blade shapes in terms of their aerodynamics [5]. In that case surrogate models usually take blade design variables as input [6, 7], but it is not mandatory, and 3D blades shape (e.g., represented as a mesh) may be passed directly to a surrogate model, if its architecture is tailored to such inputs, as in [8]. It is up to the optimizer to decide on when to use a surrogate model and when to run the whole pipeline.

The first step of the pipeline usually involves calling the API of a Computer-Aided Design (CAD) system or some other shape generation software (e.g., a generative model as described in [9]) to produce a 3D shape of the blade out of the passed design variables values. Then, follows the meshing step required for further physics calculations like CFD, CHT and/or FEA. Every step may have some post-processing routine to adjust, validate and/or extract results relevant to a specific task.

In practice every step in the design optimization cycle is done using specialized software. This leads to the fact that all steps may have different operating system (OS) requirements, dependencies for third-party libraries and execution environments needed to run the software. The multitude of various technologies involved poses a certain challenge to achieve seamless integration of the steps and effective interaction of the systems inside and outside of the optimization cycle. Also, it is often desirable to avoid vendor lock-in and support proper deployment of the resulting system so that engineers can easily utilize powerful cloud-compute environments to solve optimization tasks submitted from regular PCs or laptops.

The above-mentioned characteristics are often missing in solutions described in literature. For example, a surrogate-based integrated framework by A. Benaouali and S. Kachel [10] is based on SIEMENS NX, ICEM CFD and ANSYS FLUENT software and tailored to corresponding data formats. That makes it tightly coupled to the vendors of those systems and hard to setup, as any potential user would have to install all the required software on their PC together with GRIP and Tcl/Tk tools that are used in the framework for steps integration. In the updated solution [11], the authors enriched the framework's possibilities by supporting multidisciplinary optimization, but that resulted in an even wider set of software to install in order to use that framework, so that all the scripts that do the integration and automation work can run (e.g., a user would need to install MATLAB just to make fluid structure interpolation).

[1] What is Surrogate Model. Available: https://www.deepchecks.com/glossary/surrogate-model (Accessed 18.12.2024)

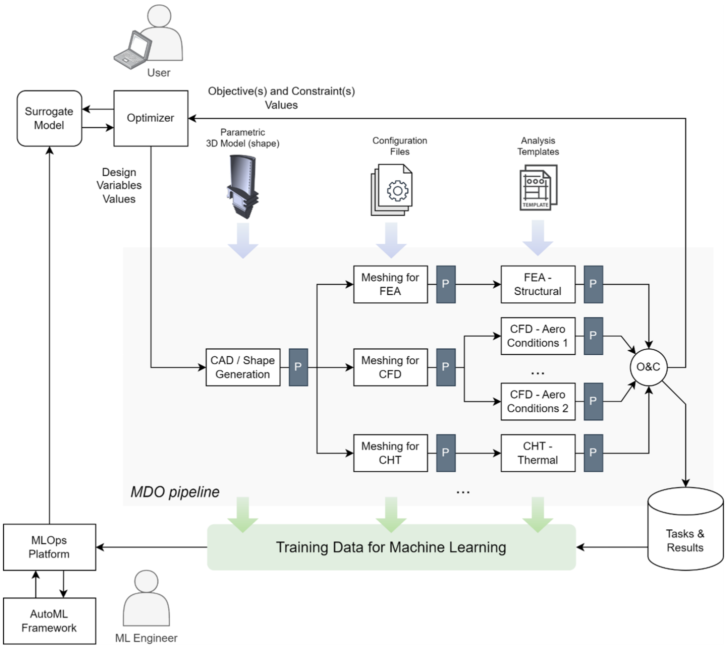


Fig. 1. Simplified workflow of MDO of a gas turbine blade

Another existing solution for surrogate-based design optimization, DADOS [12], is cloud-based, which makes it much easier to use, but on the other hand, it is not an integrated solution: DADOS automates creation and use of surrogate models but does not provide a full optimization cycle with CFD and/or other physics computations. It is expected from a user to upload the computation results to the web-interface of DADOS via Excel files. This fact significantly limits the applicability of the overall solution, because it involves manual steps, and then, surrogate models cannot be run in parallel or within the optimization cycle that involves resource-intensive tasks like CFD.

Even industrial-grade solutions, like HEEDS[2], while having a lot of support for integration with other products, still imply they all are installed on a user's PC, which makes it challenging to run a complex optimization pipeline, as it requires a user, who is typically a mechanical engineer, not an IT specialist, to setup a lot of software and custom scripts with all their dependencies that may have conflicts during installation, may be incompatible with the user's OS etc. This makes creating an integrated solution a challenging task, and while some researchers [5] have successfully built surrogate-based optimization pipelines using HEEDS and their own ML-models, that software is neither portable nor flexible enough for usage in scenarios and environments other than those described in the original paper [5], and does not support cloud deployments. A way to solve these problems is presented in this paper.

Another important part of the surrogate-based design optimization, that is missing in the known solutions, is proper lifecycle management for custom ML models, designed by an ML engineer and/or some AutoML framework. It includes a model registry and datasets storage that support versioning, linked to experiments results and history, as well as serving the learnt models in containers and/or on "as a service" basis. Lack of these characteristics forms technical debt for ML-based systems that lead to big maintenance costs [13]. In response to this problem, the MLOps paradigm has recently emerged. As stated in [14], incorporating MLOps practices is essential for bringing any ML-based solution to a production-grade quality level.

Last but not least, to avoid vendor lock-in and modern challenges of using foreign commercial software in Russia, it is important to make use of open-source and free technologies and software as much as possible, when designing new software solutions, as well as domestic providers for cloud services.

[2] Simcenter HEEDS. Available: https://plm.sw.siemens.com/en-US/simcenter/integration-solutions/heeds (Accessed 25.12.2024)

Methods

Containerization and Clouds

One of the key aspects of the suggested solution is extensive use of containers. A container is a unit of software that packages application code together with all required dependencies (libraries, execution environment etc.) and is managed by a container engine[3]. That engine, like a hypervisor for virtual machines (VMs), isolates containerized applications from the host OS, enabling the portability to run them across various infrastructures that support containers, including clouds and regular PCs[4]. The key difference between a container and a VM is that the former does not start a complete OS on top of the host system, but shares a kernel with it while keeping isolation in the user space[5] [15]. In addition, containers can be restricted in resources allocated to them, which is mainly achieved by using the cgroup mechanism[6]. The lightweight nature of containers, i.e., smaller infrastructural footprint and faster start-up times (compared to VMs) significantly contributes to cost reduction, especially when using cloud resources[7], and improves developer experience and productivity[8] [16]. Containers are widely used in ML and AI companies[9], as well as in various scientific applications, from software engineering research [17–19] to bioinformatics, where containers have been successfully applied to build infrastructure-agnostic human genome sequencing pipelines [20].

In recent years, a new cloud technology has emerged, known as serverless containers[10]. These can be considered as a more powerful kind of cloud lambda functions (that are usually considered, when referring to serverless computing), where a cloud provider accepts a complete Docker image to run instead of just source code to be executed in a selected environment. Still, no server provisioning is required from the user, and all the machine resources needed to run containerized applications are allocated on demand by the cloud service provider on a pay-as-you-go pricing model[11]. In Google's report on the state of DevOps[12], it is highlighted that cloud computing improves productivity of engineers in IT companies. This is valid for serverless containers as well, because these are as easy-to-use as serverless cloud functions and, at the same time, provide much more flexibility as regular containers[13].

In the context of this paper, the most important benefits of containers are portability, environment isolation, simplified deployment and suitability for serverless cloud operation. These characteristics are crucial to build a system that relies on a wide variety of software to run in organized and scalable pipelines launched from a laptop by users, who are engineers, but not IT experts. The portability of

[3] Understanding containers. Available: https://www.redhat.com/en/topics/containers (Accessed 25.12.2024); Susnjara S., Smalley I. What Is Containerization? Available: https://www.ibm.com/think/topics/containerization (Accessed 25.12.2024); Chto takoye konteynerizatsiya: Oblachnaya terminologiya [What is containerization: Cloud terminology]. Available: https://yandex.cloud/ru/docs/glossary/containerization (Accessed 24.12.2024)

[4] Susnjara S., Smalley I. What Is Containerization? Available: https://www.ibm.com/think/topics/containerization (Accessed 25.12.2024); Chto takoye konteynerizatsiya: Oblachnaya terminologiya [What is containerization: Cloud terminology]. Available: https://yandex.cloud/ru/docs/glossary/containerization (Accessed 24.12.2024)

[5] Susnjara S., Smalley I. What Is Containerization? Available: https://www.ibm.com/think/topics/containerization (Accessed 25.12.2024)

[6] Menage P., Lameter C., Jackson P. Control Groups. Available: https://docs.kernel.org/admin-guide/cgroup-v1/cgroups.html (Accessed 25.12.2024)

[7] Susnjara S., Smalley I. What Is Containerization? Available: https://www.ibm.com/think/topics/containerization (Accessed 25.12.2024); Chto takoye konteynerizatsiya: Oblachnaya terminologiya [What is containerization: Cloud terminology]. Available: https://yandex.cloud/ru/docs/glossary/containerization (Accessed 24.12.2024); The Total Economic Impact™ of Docker Business. Available: https://www.docker.com/resources/tei-of-docker-business-a-conversation-with-our-cro-webinar/ (Accessed 16.05.2025)

[8] The Total Economic Impact™ of Docker Business. Available: https://www.docker.com/resources/tei-of-docker-business-a-conversation-with-our-cro-webinar/ (Accessed 16.05.2025); Hayzen A. How containers improve the way we develop software. Available: https://www.embedded.com/how-containers-improve-the-way-we-develop-software (Accessed 25.12.2024); Choroomi A. How Kinsta Improved the End-to-End Development Experience by Dockerizing Every Step of the Production Cycle | Docker. Available: https://www.docker.com/blog/how-kinsta-improved-the-end-to-end-development-experience-by-dockerizing-every-step-of-the-production-cycle (Accessed 25.12.2024)

[9] Hype Cycle for Container Technology, 2024. Available: https://www.gartner.com/en/documents/5521795 (Accessed 16.05.2025)

[10] Susnjara S., Smalley I. What Is Containerization? Available: https://www.ibm.com/think/topics/containerization (Accessed 25.12.2024); Hype Cycle for Container Technology, 2024. Available: https://www.gartner.com/en/documents/5521795 (Accessed 16.05.2025); Yandex Serverless Containers. Available: https://yandex.cloud/ru/docs/serverless-containers (Accessed 25.12.2024)

[11] Susnjara S., Smalley I. What Is Containerization? Available: https://www.ibm.com/think/topics/containerization (Accessed 25.12.2024)

[12] Accelerate State of DevOps 2023. Available: https://services.google.com/fh/files/misc/2023_final_report_sodr.pdf (Accessed 16.05.2025)

[13] Susnjara S., Smalley I. What Is Containerization? Available: https://www.ibm.com/think/topics/containerization (Accessed 25.12.2024); Yandex Serverless Containers. Available: https://yandex.cloud/ru/docs/serverless-containers (Accessed 25.12.2024)

containers allows running the same Docker image in the cloud and locally, thus enabling a quick feedback loop, when developing the image (before releasing it) and flexibility to run lightweight pipelines on a regular PC (while setting up the optimization task and physics calculations templates), before rolling out a full-fledged workload to the cloud.

The Proposed Architecture

Key aspects of the proposed architecture include:

- containerization of software components,
- integration of those in pipelines in the optimization loop,
- storage for all artifacts and settings of pipelines' stages,
- software developed to enable system operation in the cloud,
- utilization of serverless containers in the cloud,
- using Docker Compose when running locally or on VMs,
- AutoMLOps for maintaining surrogate models' lifecycle.

The proposed system's serverless operation mode requires the cloud environment to support the well-known APIs for Simple Storage Service (S3), Simple Queue Service (SQS) and DynamoDB in document mode, as well as the ability to store Docker images and start containers by triggers from SQS. This set was originally introduced by Amazon Web Services (AWS), but now it is available both in foreign cloud providers (Amazon Web Services, Google Cloud Compute, Microsoft Azure) and in domestic solutions, such as Yandex Cloud[14], as it is a de facto standard in this industry. The serverless mode of operation helps to achieve high scalability of the system, provides separation of computing resources and data storage, and gives the opportunity to use built-in monitoring and information security tools provided by the cloud offering.

Fig. 2 shows the general scheme of the proposed system architecture, using the Amazon Web Services notation.

A more detailed structural scheme of the pipeline with indication of used technologies is shown in Fig. 3. In this case, the components are mapped to Yandex Cloud services, since the system was implemented using the resources of this cloud provider.

The system uses Dakota[15] as an optimizer, an open-source solution that includes many optimization algorithms, including those that support global optimization using surrogate models. In addition to optimization, Dakota has a design of experiments (DoE) functionality, which allows sampling of blades parameters to achieve uniform coverage of the search space (e.g., using the Latin hypercube algorithm), which is useful for training surrogate ML-models.

At each iteration in the optimization process, Dakota invokes the Cloud Task Runner program, which is written by the authors in the Go language. This program is responsible for generating a task and starting one run of the pipeline. It writes to the database the information necessary to launch the pipeline, including the values of variables describing the blade and configuration files for the components, as well as the sequence of their execution. Each component is responsible for its own stage. The hierarchy "task – run – stage" is reflected in the structure of S3 storage (Fig. 3), which is used to store all artifacts produced by the components of the pipeline, and for data exchange between them (together with SQS messaging). The artifacts are divided into four groups: configuration files, input files, output files and additional files. The need for particular files is determined by each specific component. An example of additional files is an archive with complete CFD calculation results, which often has a size of hundreds of megabytes or more depending on the mesh detail and the number of steps in the calculation. A cheaper class of S3 storage (standard infrequent access) is used for additional files.

Each component is deployed as a Docker container, which includes the component's software and a program developed by the authors called Cloud Connector. This program is written in the Go language

[14] Yandex Serverless Containers. Available: https://yandex.cloud/ru/docs/serverless-containers (Accessed 25.12.2024)

[15] Dakota. Available: https://dakota.sandia.gov (Accessed 11.01.2025)

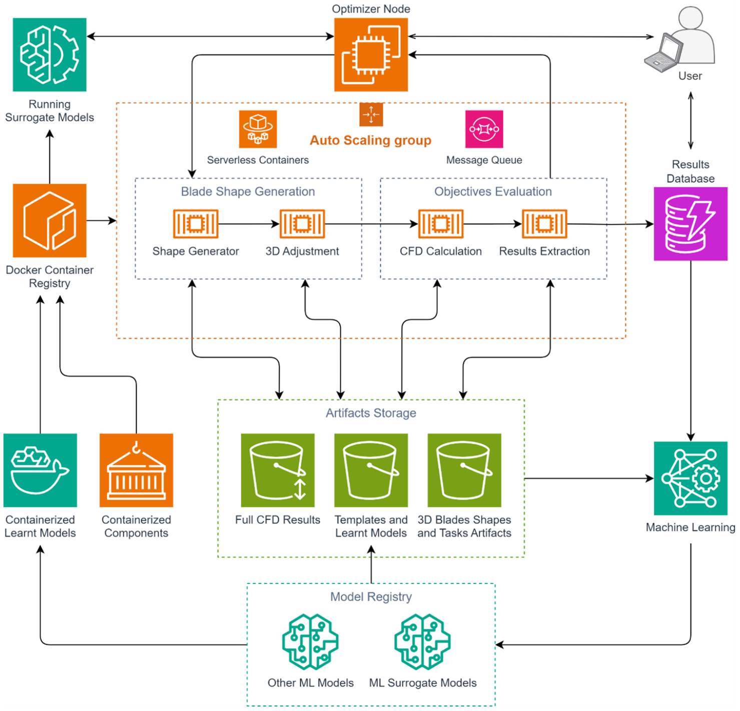


Fig. 2. The proposed system's architecture for cloud deployment

and is responsible for universal interaction of components with cloud services: SQS, S3, DynamoDB – through the authors' library Cloud Task Registry, that provides an application programming interface Task Registry API and is used both in the Cloud Connector and Cloud Task Runner programs. Together these programs and the library make up the software complex named Cloud Optimization Suite.

After solving an optimization problem formulated by the user in the Dakota interface, all configuration, input and output files and other artifacts for each of the stages of each run remain recorded in the S3 storage. This includes, among other things, 3D shapes of the generated blades and complete CFD calculation results. This makes each step fully reproducible: a user can download the relevant files, run the component responsible for the stage on a local machine and study its operation in detail for specific parameters or take the files of interest for further work with them. Similar functionality of data storage is implemented in the local operation mode using Docker Compose, but in this case, data is stored on disk in the directory specified by the user, and the cloud functionality is not used, which allows working with the proposed system locally and fully independent of cloud providers, or on VMs.

Training of Surrogate Models

Parameters of generated blades' shapes along with CFD results and other artifacts essentially form datasets for surrogate models training. This feature of the proposed architecture enables straightforward integration with AutoML frameworks. These allow searching for best regression models automatically. Alternatively, ML-engineers may design more sophisticated models within the MLOps framework, as was described earlier. In combination, this leads to an AutoMLOps solution. In the proposed system, ClearML open-source AI platform[16] is used for that purpose together with Auto-Sklearn 2.0 framework [21] for AutoML support. ClearML has a free version (Apache-2.0 license) and uses exclusively open-source components, allowing integration with a lot of other frameworks and services. Since ClearML supports integration with AWS S3, the training data can be used directly from the artifacts storage (Fig. 2).

Using the MLOps platform, the following ML pipeline is designed:

1. Load new training data from the S3 storage.

[16] ClearML Open Source Platform. Available: https://clear.ml/ai-platform (Accessed 11.01.2025)

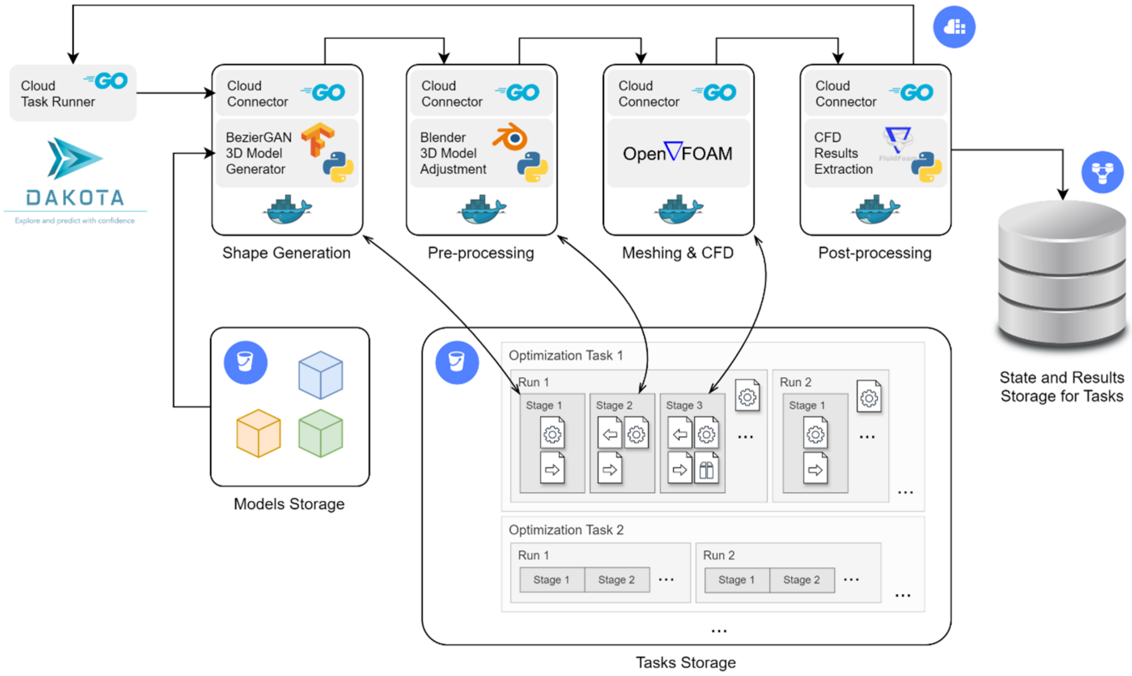


Fig. 3. Structure of the pipeline and technologies used

2. Preprocess the data if needed (e.g., to extract some extra features).
3. Call the AutoML framework for training of a surrogate model.
4. Test the model performance.
5. Optionally, perform hyperparameters tuning.
6. Store the model and assign a version to it.

Pipeline orchestration is performed by the ClearML server (Fig. 4), and the training itself is done by ClearML agents, which are deployed on machines that have all required hardware resources (CPUs and/or GPUs).

Communication with the agents is done via tasks queues. The trained models are stored in the model storage (e.g., in an S3 bucket). In order to perform inference using saved models, ClearML provides a serving solution. It allows to expose HTTP endpoints to accept inference requests and returns corresponding responses after delegating the processing to a configured serving engine (ClearML offers its own CPU-based engine and also supports Triton[17] by Nvidia for GPU-enabled hardware). In the proposed system, the surrogate model interface of Dakota is used to make such inference calls, passing blades geometric parameters that are subject to optimization.

By the point, when surrogate models training is performed, a large number of expensive calculations is typically done, and CFD results are obtained for a multitude of blade shapes that may be confidential as a company asset, as well as surrogate models themselves. Thus, we obtain a large amount of sensitive information in a single step, possibly even on a single VM, which necessitates additional protection. This topic was covered in [22–24] where the Trusted AutoML task was formalized and addressed from performance and cybersecurity perspectives, including but not limited to the application for training surrogate models on gas turbines' data.

Implementation

The Cloud Optimization Suite software is implemented using the Go programming language (Golang). The choice of the language was made to fulfill the need to minimize the footprint of the Cloud

[17] Triton Inference Server. Available: https://developer.nvidia.com/triton-inference-server (Accessed 18.01.2025)

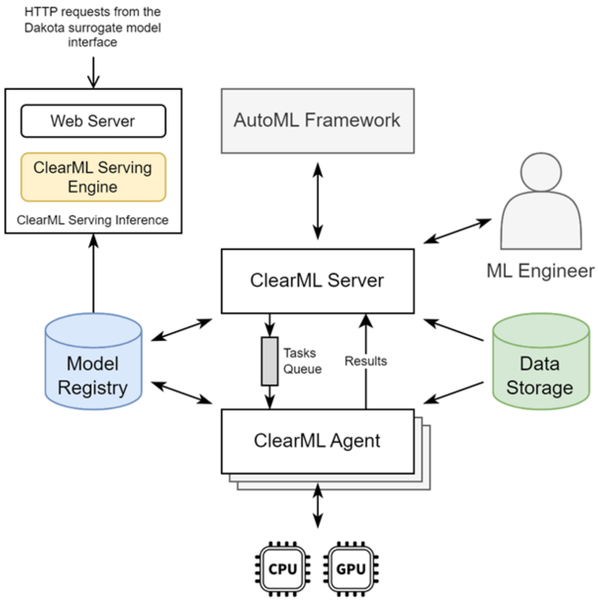


Fig. 4. Detailed view on the MLOps platform used in the system

Connector on container images of the pipeline stages components' software and simplify integration. Programs in Go are compiled into natively executable binaries and can be statically linked with all used libraries, thus making the program independent from other system libraries. As a result, installation of the Cloud Connector is just copying one executable file into a Docker container. Golang provides a garbage collector, which simplifies memory management and a set of useful concurrent programming mechanisms (goroutines, channels etc.), which are particularly helpful when dealing with asynchronous calls.

One of Golang's distinctive features is absence of an exception facility in the language (i.e., there is no control structure associated with error handling), and errors are handled in the same way as regular variables values. The language authors stand for that decision, claiming that "explicit error checking forces the programmer to think about errors – and deal with them – when they arise"[18]. Although at the moment there is no scientific evidence that software developed using Go has better quality than that written in other languages [25], explicit error handling has proven to be useful in the process of developing the Cloud Optimization Suite and improved its robustness by enforcing consideration of many exceptional scenarios beforehand.

The Cloud Optimization Suite source code is available at GitHub[19] under the Apache-2.0 permissive license and consists of three Go modules: Cloud Task Registry, Cloud Connector and Cloud Task Runner. The first one is a library for communication with cloud resources and two others are applications: Cloud Task Runner is placed at the machine, where the Dakota optimizer is installed, and Cloud Connector is appended to each components' Docker image and used as an entry point (with additional arguments, like a stage name, the main component executable path, DynamoDB document API URL etc., provided either at Docker build time or via environment variables).

Cloud Task Runner populates the database tables with a new record for a task run and corresponding records for its stages according to a configuration file and then submits the task for processing. It also copies configuration files for the stages' components to an S3 bucket according to the structure shown in Fig. 3. After submitting the task for execution, the Cloud Task Runner program waits for the task run completion via long polling of the final SQS queue, while the task run's Universally Unique Identifier

[18] Pike R. Go at Google: Language Design in the Service of Software Engineering. Available: https://go.dev/talks/2012/splash.article (Accessed 09.02.2025)

[19] Cloud Optimization Suite. Available: https://github.com/wndrws/cloud-optimization-suite (Accessed 08.02.2025)

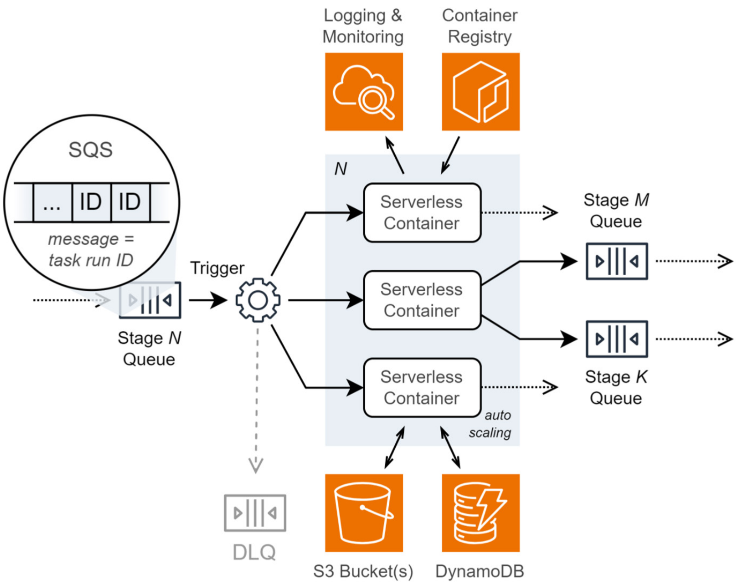


Fig. 5. Functional diagram of a stage within the pipeline

(UUID) makes its way through the pipeline. The generic functional diagram of a pipeline stage is given in Fig. 5. Arrows with big heads represent message flow, and arrows with smaller heads stand for data flow (Docker images, files, database communication, logs etc.).

Each pipeline stage is associated with one software component that implements the stage's functionality and deployed as a set of serverless container instances. There may be none, one or many container instances at each stage in any given moment in time depending on the number of messages in processing and scaling configuration in the cloud console. Yandex Cloud uses triggers to monitor SQS queues, and when a new message arrives, it creates a new container instance and passes the message to it; then, it shuts down the instance if all messages have been processed and the queue is empty. When a serverless container instance finishes a task run processing, it uploads output files to S3 (using the Cloud Connector) and sends the run's UUID to the SQS queues of the next adjacent stage(s). All logs written by the component are collected by the cloud provider and managed using the Cloud Logging service. Thus, the system administrator can see logs from all containers in one place in the cloud console together with monitoring information, like CPU load and RAM consumption of the software in serverless containers.

When a message from an SQS queue is passed to a serverless container, it goes to a Cloud Connector's handler. The algorithm of the Cloud Connector program can be briefly formulated as follows (for each stage):

1. Cloud Connector waits for messages from the queue in SQS. The incoming message contains a run UUID and initiates the start of the stage.

2. After requesting information about the task and the stage from DynamoDB, Cloud Connector downloads the configuration and input files of the stage from S3.

3. Using the configured component start command, Cloud Connector creates the corresponding subprocess and waits for it to complete. In parallel, Cloud Connector makes periodic queries to DynamoDB, checking the status of the task to see if it has been canceled. If the task has been canceled, a termination signal (SIGTERM) is sent to the subprocess.

4. Based on the value of the subprocess return code, Cloud Connector determines the fact of successful or unsuccessful execution of the pipeline stage: a non-zero return code means that an error occurred.

5. If the component startup was successful, Cloud Connector reads the output file (if the path was specified) and loads it into S3. A message with the run ID is then sent to the SQS queue(s) defined for the stage(s) immediately following the current stage.

6. If this stage is the last stage in the run (i.e., has no next stages in the task pipeline configuration), the output file is read as an associative array of objectives (and/or constraints) and their values and written as the task run results to DynamoDB.

7. If paths to additional artifacts are configured for the step, these folders and/or files are compressed into an archive and uploaded to S3.

If an error occurs during a Cloud Connector operation, the corresponding status is set for this stage of the task, the Cloud Connector terminates, and then the container is restarted by means of the cloud provider. If the limit of restarts is exhausted, the task run ID is sent to a special Dead Letter Queue, which is not related to the pipeline stages, but can be used for debugging.

After the last step of the run is completed, its identifier is sent to a special SQS queue, from which it is read by the Cloud Task Runner program. Then, it outputs a report listing all the attributes, status and execution time of each stage and the run as a whole. The cycle repeats until a convergence condition (configured in the optimizer) is reached or the task is cancelled.

The data model of DynamoDB tables is shown in Tables 1 and 2: one table is used to store information about task runs, and the other one is for their stages at each run.

Table 1

DynamoDB data model – the task runs table

Keys	Name	Type	Explanation
PK	task_id	string	Identifier of the task
SK, GSI PK	run_uuid	string	Identifier of the task run
	parameters	map<string, number>	Design variables values for the run
	results	map<string, number>	Objectives and constraints values
	task_definition	string	The task configuration from Dakota
	creation_time	number	Time when the task was created
	status	string	Status of the task run

There is a one-to-many relationship between the stored entities: each task run record is associated with many task stage records. Saving timestamps of start and completion of each stage for every task run is useful for monitoring and collecting statistics of the system functioning. The tables are created automatically, when Cloud Task Registry is used with an empty DynamoDB (by the means of migration code implemented in Go).

The DynamoDB keys abbreviations are as follows: PK – partition key, SK – sort key, GSI – global secondary index. Possible statuses in the task runs table are Pending, InProgress, Success, Error, Cancelled, – and possible statuses in the task stages table are Submitted, Finished, Failed and Cancelled.

It is worth noting that the system provides users with an ability to cancel tasks execution, forcing the running containers to interrupt component's software, do the required cleanup and state management and terminate.

Results and Discussion

After implementation, the system was tested in a series of experiments, using Yandex Cloud resources, as well as local Docker Compose deployments. The experiments performed can be grouped as follows:

1. Testing the system operation in the DoE mode with different numbers of parallel running pipelines to assess the system's scalability and correctness of its functioning in both variants of deployment.
2. Checking the system operation in optimization mode to find the gas turbine blade shape that maximizes the aerodynamic efficiency coefficient with and without the use of surrogate models, comparison of the found optima and the time spent.
3. Validation of correctness of CFD results produced using the system from the domain perspective.

Table 2

DynamoDB data model – the task stages table

Keys	Name	Type	Explanation
PK, GSI PK	run_uuid	string	Identifier of the task run
SK	n_ord	number	Ordinal number of the stage
GSI SK	name	string	Name of the stage
	status	string	Status of the task run at this stage
	config	string	Paths to configuration, input and output files of the component at this stage in the S3 bucket (see below)
	input	string	
	output	string	
	t_start_utc	number	Times when the task run processing started and finished at this stage
	t_finish_utc	number	
	executor	string	Execution environment information (e.g., revision of the Docker image used for the run)
	s3_bucket	string	Name of the S3 bucket used
	comment	string	Any additional information from the component
	next	list<string>	List of the adjacent next stages names

The obtained results have shown that the developed system functions correctly both in a single-threaded environment and when running pipelines in parallel (in both deployment modes: serverless and on a single machine via Docker Compose); the integrity of data accessed concurrently is not violated; the system scales vertically according to the computational resources of the processor and horizontally according to the number of available physical cores (Table 3): Hyper-Threading technology does not give advantages, which was revealed during experiments on the Intel Core i7-7700K CPU. When used in serverless mode, horizontal scaling is done by utilizing more instances of serverless containers in parallel within quotas set by the cloud service provider.

Table 3

Time spent on DoE using the suggested system on a PC with 4 physical and 8 logical CPU cores

Concurrent pipelines count	1	2	3	4	5	6	7	8
Total time, s	15970	8655	6617	5830	5840	5713	5894	5730

The use of surrogate models in the Efficient Global Optimization mode led to at least 2.5 times reduction in time for the optimal blade shape search compared to optimization without surrogate models, given the same number of the pipeline runs. The achieved optimal value of the aerodynamic efficiency coefficient was different only by 2.46% in favor of the optimization without surrogate models. In serverless mode the time spent on optimization (in fixed conditions) was 18–36% less compared to running on a VM with the same number of processor cores of the same architecture (Intel Haswell) and the same

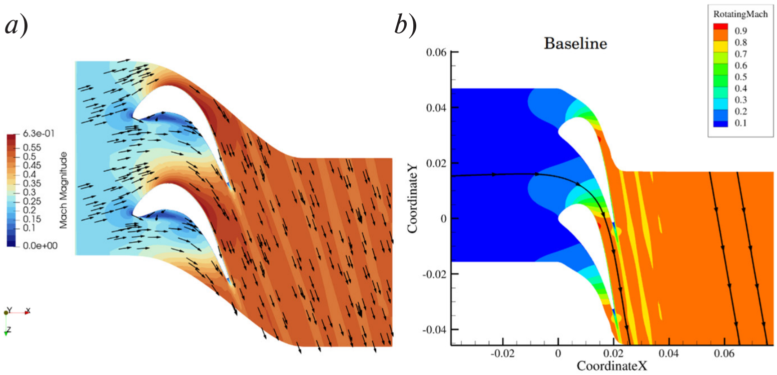


Fig. 6. The velocity field obtained by the authors (*a*) and a reference (*b*) from [26]

clock frequency of 2.1 GHz (2 GB RAM per core allocated by the cloud provider was more than enough for all containerized components).

Validation of CFD results produced using the system was successful according to the set of quantitative checks and qualitative analysis performed by the authors. For example, Fig. 6 shows a comparison of the velocity field (in Mach numbers) obtained by the authors and a similar reference field from [26].

A separate discussion should be given to the architectural advantages of the system. As its components communicate asynchronously via message queues (SQS), which transport only task runs' identifiers, loose coupling is achieved, providing the following benefits:

Fault-tolerance and resilience. If an error occurs during functioning of a pipeline component, the task run processing request returns to the queue and becomes visible for other containers at the stage[20] that can handle the processing request, while the failed serverless container is shut down by the platform and replaced by a new instance if needed for serving the following requests. The recover process does not impact the functioning of other system parts, e.g., parallel pipeline runs continue unaffected, though the erroneous task status change is reflected in the DynamoDB and is visible via serverless containers monitoring. Status tracking of individual task runs also protects the system from using any partly written data that may be left by failed serverless containers. In case a container becomes unresponsive, it also will be terminated by the cloud platform according to the configured timeout (separately for each pipeline stage).

Hot swap of Docker images. When the root-cause of an error is identified in a containerized component and fixed, it is possible to upload the new Docker image and replace the one currently utilized by the serverless container without stopping the pipeline operation (Yandex Cloud gracefully stops the running container instance and starts a new one instead). The same task runs, that were not passing previously, will then be automatically retried using the latest component version[21]. This hot swap support is highly beneficial for lengthy and costly optimization tasks, when restarting from the beginning in case of any error is not affordable.

Flexibility in data exchange formats and software selection. Since in the proposed architecture there are no requirements for data formats used by pipelines components, they can be chosen freely by the implementations and must be agreed only between the neighboring stages (as one stage's output files usually serve as the next stage's inputs). This, together with containerization, enables flexibility in selecting the formats and software used at each pipeline stage, without the need to make any adjustments to the system to maintain compatibility. By abstaining high-level code (i.e., the Cloud Optimization Suite) and the system operation principles from low-level details, the architecture follows the dependency inversion principle [27] and the open-closed principle [28] widely adopted in software engineering [29].

[20] Trigger for Message Queue that sends messages to the Serverless Containers container. Available: https://yandex.cloud/en-ru/docs/serverless-containers/concepts/trigger/ymq-trigger (Accessed 04.02.2025)

[21] Yandex Serverless Containers. Available: https://yandex.cloud/ru/docs/serverless-containers (Accessed 25.12.2024)

Table 4

Comparison of the suggested system and existing solutions

Criteria	Ours	Song et. al., 2023 (DADOS) [12]	Benaouali and Kachel, 2017 [10]	Benaouali and Kachel, 2019 [11]
Integrated solution	Yes	No	Yes	Yes
Data file formats	Arbitrary	Excel	Parasolid, FluentMesh	Parasolid, IGES, FluentMesh, PATRAN, NASTRAN
Software components in the pipeline	Any containerizable (using Docker)	Any (at the user's side)	Siemens NX, ICEM CFD, FLUENT	Siemens NX, ICEM CFD, FLUENT, MSC.PATRAN, MSC.NASTRAN, MATLAB
Environment	Any cloud providing S3, DynamoDB and SQS APIs or Docker Compose	Cloud (China)	Local	Local
Supported surrogate models	Any models provided by Dakota or own ML-models via ClearML MLOps	PRS, RBF, KRG, MLS, SVR, and ANN (feed-forward)	Gaussian RBF	RBF
Optimization algorithms	DIRECT, EGO and others supported by Dakota[22]	In-house + SA, swarm, genetic and other algorithms	Sequential Quadratic Programming	Genetic Algorithm
Supported DoE sampling	LHS, Monte-Carlo, Rank-1 Lattice, Digital Nets (Sobol)	OA, CCD, LHS, OLHS	LHS	Improved LHS
Parallel steps in pipelines	Yes	No	No	Yes
Multi-disciplinary	Yes	No	No	Yes
Publicly available	Yes	Yes (after registration)	No	No
Open source	Yes	No	No	No

Scalability. Using asynchronous communication and cloud services, plus separation of storage and compute makes the system highly scalable. New instances of serverless containers are created automatically by the cloud platform in response to new messages in task queues. As soon as some task runs are finished, unnecessary containers are shut down and don't incur any more costs in accordance with the pay-as-you-go model. More parallel pipelines or parallel branches within one pipeline do not induce resource contention, since the DynamoDB database, SQS queues and S3 buckets (that are used for data exchange and storage), are designed for efficient handling of concurrent loads, and apart from them there are no shared resources that containers can contend for. Vertical scaling is also supported, as serverless containers are configurable in terms of CPU cores and RAM allocated by the cloud provider[23]. Finally, as the system compute resources scale down to zero in absence of tasks, there is no risk of wasting resources due to human factor, as it can be with VMs, which are sometimes unintentionally left powered on for a weekend without any workload, producing unwanted spendings.

[22] Optimization Usage Guidelines. Available: https://snl-dakota.github.io/docs/6.21.0/users/usingdakota/studytypes/optimization.html#opt-usage (Accessed 10.02.2025)

[23] Runtime environment. Available: https://yandex.cloud/en-ru/docs/serverless-containers/concepts/runtime (Accessed 04.02.2025)

Apart from the abovementioned advantages, the system administrator can also make use of cloud services that come together with the Serverless Containers service: monitoring (including performance charts and logging), access and secrets management, billing details etc.

To compare the suggested system with the existing solutions discussed previously, a set of criteria was devised considering the most relevant characteristics within the context of this research, and the comparison results are presented in Table 4. Used abbreviations: PRS – polynomial response surface, RBF – radial basis functions, KRG – kriging, MLS – moving least squares, SVR – support vector regression, ANN – artificial neural network; DIRECT – dividing rectangles, EGO – efficient global optimization, SA – simulated annealing; LHS – Latin hypercube sampling, OLHS – optimal LHS, CCD – central composite design, OA – orthogonal arrays.

From that table it becomes even more apparent that the suggested system is more generic and gives more possibilities to build engineering optimization pipelines from various components without the need to install them on every user's PC, plus the support for ML-models that can use full power of modern ML frameworks, AutoML and MLOps for proper models' lifecycle management.

Last but not least, the system does not rely on any closed-source components and avoids cloud vendor lock-in as far as possible by sticking to the most common set of cloud services APIs, namely the S3, SQS and DynamoDB in document mode, which are available in many Russian and foreign cloud providers. All the open-source and free software used have enterprise-friendly licensing (i.e., no copyleft), which is crucial for open-source software clearing purposes and enables commercial use of the system.

Conclusions

In this paper a computer system architecture was proposed together with its implementing software for rapid prototyping of gas turbine blades, which allows prediction of their physical properties by given geometric parameters. This is an integrated cloud-based solution, though flexible enough to be run locally on a regular PC, without the need for complex environment setup for users (mechanical engineers). Despite the fact that the system was described in application to gas turbine blades aerodynamics optimization, it is not restricted to that domain and can be readily used in design optimization tasks for other kinds of industrial objects and physics disciplines.

The next steps in this research direction are to support the hybrid mode of operation, in which the system will utilize serverless containers simultaneously with dedicated VMs for long-running evenly distributed workloads, and to support non-parametric surrogate models that take 3D blade geometry as input [8, 30] to investigate their performance and generalization potential.

REFERENCES

1. **Hammond J., Pepper N., Montomoli F., Michelassi V.** Machine learning methods in CFD for turbomachinery: A review. *International Journal of Turbomachinery, Propulsion and Power*, 2022, Vol. 7, No. 2, Art no. 16. DOI: 10.3390/ijtpp7020016

2. **Martins J.R.R.A., Ning A.** *Engineering Design Optimization*. Cambridge, UK: Cambridge University Press, 2022. DOI: 10.1017/9781108980647

3. **Jiang P., Zhou Q., Shao X.** *Surrogate Model-Based Engineering Design and Optimization*. Singapore: Springer, 2020. DOI: 10.1007/978-981-15-0731-1

4. **Xu L., Jin S., Ye W., Li Y., Gao J.** A review of machine learning methods in turbine cooling optimization. *Energies*, 2024, Vol. 17, No. 13, Art. no. 3177. DOI: 10.3390/en17133177

5. **Zhang C., Janeway M.** Optimization of turbine blade aerodynamic designs using CFD and neural network models. *International Journal of Turbomachinery, Propulsion and Power*, 2022, Vol. 7, No. 3, Art. no. 20. DOI: 10.3390/ijtpp7030020

6. **Kulfan B.M.** Universal parametric geometry representation method. *Journal of Aircraft*. 2008, Vol. 45, No. 1, Pp. 142–158. DOI: 10.2514/1.29958

7. **Zhemelev G.** Parameterized 3D representation of gas turbine blades for machine learning applications. *2024 Wave Electronics and its Application in Information and Telecommunication Systems* (*WECONF*), 2024, Pp. 1–6. DOI: 10.1109/WECONF61770.2024.10564621

8. **Mou S., Bu K., Ren S., Liu J., Zhao H., Li Z.** Digital twin modeling for stress prediction of single-crystal turbine blades based on graph convolutional network. *Journal of Manufacturing Processes*, 2024, Vol. 116, Pp. 210–223. DOI: 10.1016/j.jmapro.2024.02.054

9. **Zhemelev G.A.** Automatic synthesis of 3D gas turbine blades shapes using machine learning. *Information Security Problems. Computer Systems*, 2024, Vol. 59, No. 2, Pp. 152–168. DOI: 10.48612/jisp/dx8x-2he5-tffd

10. **Benaouali A., Kachel S.** A surrogate-based integrated framework for the aerodynamic design optimization of a subsonic wing planform shape. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 2017, Vol. 232, No. 5, Pp. 872–883. DOI: 10.1177/0954410017699007

11. **Benaouali A., Kachel S.** Multidisciplinary design optimization of aircraft wing using commercial software integration. *Aerospace Science and Technology*, 2019, Vol. 92, Pp. 766–776. DOI: 10.1016/j.ast.2019.06.040

12. **Song X., Wang S., Zhao Y., Liu Y., Li K.** DADOS: A cloud-based data-driven design optimization system. *Chinese Journal of Mechanical Engineering* (*English Edition*), 2023, Vol. 36, Art. no. 34. DOI: 10.1186/s10033-023-00857-x

13. **Sculley D., Holt G., Golovin D., Davydov E., Phillips T. et al.** Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 2015, Vol. 28, pp. 2503–2511.

14. **Kreuzberger D., Kühl N., Hirschl S.** Machine learning operations (MLOps): Overview, definition, and architecture. *IEEE Access*, 2023, Vol. 11, Pp. 31866–31879. DOI: 10.1109/ACCESS.2023.3262138

15. **Bentaleb O., Belloum A.S.Z., Sebaa A., El-Maouhab A.** Containerization technologies: taxonomies, applications and challenges. *The Journal of Supercomputing*, 2022, Vol. 78, No. 1, Pp. 1144–1181. DOI: 10.1007/s11227-021-03914-1

16. **Koskinen M., Mikkonen T., Abrahamsson P.** Containers in software development: A systematic mapping study. *Product-Focused Software Process Improvement* (*PROFES 2019*), 2019, Vol. 11915, Pp. 176–191. DOI: 10.1007/978-3-030-35333-9_13

17. **Kim B.S., Lee S.H., Lee Y.R., Park Y.H., Jeong J.** Design and Implementation of cloud docker application architecture based on machine learning in container management for smart manufacturing. *Applied Sciences*, 2022, Vol. 12, No. 13, Art. no. 6737. DOI: 10.3390/app12136737

18. **Bobunov A.** Using containerization to simplify and accelerate testing processes in financial organizations. *International Journal of Humanities and Natural Sciences*, 2024, Vol. 95, No. 8–1, Pp. 113–117. DOI: 10.24412/2500-1000-2024-8-1-113-117

19. **Cito J., Ferme V., Gall H.C.** Using Docker containers to improve reproducibility in software and web engineering research. In: *Web Engineering* (eds. A. Bozzon, P. Cudre-Maroux, C. Pautasso), 2016, Vol. 9671, Pp. 609–612. DOI: 10.1007/978-3-319-38791-8_58

20. **Kadri S., Sboner A., Sigaras A., Roy S.** Containers in bioinformatics: Applications, practical considerations, and best practices in molecular pathology. *Journal of Molecular Diagnostics*, 2022, Vol. 24, No. 5, Pp. 442–454. DOI: 10.1016/j.jmoldx.2022.01.006

21. **Feurer M., Eggensperger K., Falkner S., Lindauer M., Hutter F.** Auto-Sklearn 2.0: Hands-free AutoML via Meta-Learning. *arXiv:2007.04074*, 2020. DOI: 10.48550/arXiv.2007.04074

22. **Bezzateev S.V., Fomicheva S.G., Zhemelev G.A.** Trusted automatic machine learning in the operation of digital twins. *T-Comm*. 2024, Vol. 18, No. 7, Pp. 44–55. DOI: 10.36724/2072-8735-2024-18-7-44-55

23. **Bezzateev S.V., Fomicheva S.G., Zhemelev G.A.** Techniques for accelerating algebraic operations in agent-based information security systems. *2023 Wave Electronics and its Application in Information and Telecommunication Systems* (*WECONF*), 2023, Pp. 1–6. DOI: 10.1109/WECONF57201.2023.10147978

24. **Bezzateev S.V., Zhemelev G.A., Fomicheva S.G.** Research on the performance of AutoML platforms under confidential computing. *Information Security Problems. Computer Systems*, 2024, Vol. 61, No. 3, Pp. 109–126. DOI: 10.48612/jisp/abff-du38-v739

25. **Ray B., Posnett D., Devanbu P., Filkov V.** A large-scale study of programming languages and code quality in GitHub. Communications of the ACM, 2017, Vol. 60, No. 10, Pp. 91–100. DOI: 10.1145/3126905

26. **Aissa M.H.** GPU-accelerated CFD Simulations for Turbomachinery Design Optimization. Delft, Netherlands: Delft University of Technology, 2018. DOI:10.4233/UUID:1FCC6AB4-DAF5-416D-819A-2A7B0594C369

27. **Martin R.C.** OO design quality metrics: An analysis of dependencies. Report on Object Analysis and Design, 1995, Vol. 2.

28. **Meyer B.** Object-oriented Software Construction (Prentice-Hall International series in computer science). NJ: Prentice-Hall, 1988.

29. **Martin R.C.** Design principles and design patterns. Object Mentor SOLID Design Papers Series, 2000, Vol. 1, No. 1, Pp. 1–34.

30. **Cao J., Li Q., Xu L., Yang R., Dai Y.** Non-parametric surrogate model method based on machine learning with application on low-pressure steam turbine exhaust system. Journal of the Global Power and Propulsion Society, 2022, Vol. 6, Pp. 165–180. DOI: 10.33737/jgpps/151661

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Zhemelev Georgiy A.
Жемелев Георгий Алексеевич
E-mail: wws.dev@gmail.com
ORCID: https://orcid.org/0000-0001-7126-6787

Drobintsev Pavel D.
Дробинцев Павел Дмитриевич
E-mail: drob@ics2.ecd.spbstu.ru
ORCID: https://orcid.org/0000-0003-1116-7765

Submitted: 19.02.2025; Approved: 17.04.2025; Accepted: 29.04.2025.

Поступила: 19.02.2025; Одобрена: 17.04.2025; Принята: 29.04.2025.