Стохастическое моделирование работы системы автоматической обработки информации

Т.М. Попова

Тихоокеанский государственный университет, Хабаровск

Аннотация: В работе рассмотрена стохастическая модель работы системы автоматической обработки информации, которая описываемая системой дифференциальных уравнений Колмогорова распределения вероятностей состояний, в предположении, что поток заявок пуассоновский, в том числе простейший. Предложена схема решения системы дифференциальных уравнений большой размерности с мало меняющимися начальными данными, проведено сравнение параметров представленной модели с параметрами имитационной модели работы веб-сервера Apache HTTP Server. Для сравнения имитационной и стохастической моделей между собой использован тестовый сервер с генераций запросов и имитацией их обработки с использованием статистики программой Apache JMeter, на основании которого оценены значения параметров интенсивностей потоков входящих и обработанных заявок. Представленная модель не противоречит имитационной, позволяет оценить состояния системы при различных режимах работы и рассчитать загруженности веб-сервера при большом потоке данных.
Ключевые слова: стохастическое моделирование, имитационная модель, уравнения Колмогорова, метод прогонки, система массового обслуживания, характеристики эффективности, тестовый сервер, поток заявок, каналы обслуживания, очередь.

В условиях экспоненциального роста интернет-трафика и увеличения сложности веб-приложений, обеспечение устойчивой работы серверных инфраструктур требует современных подходов к анализу и прогнозированию. Любая вычислительная система или система обработки данных находится под влиянием случайных факторов, влияющих на режим его работы, следовательно, представляет собой случайный процесс, обусловленный случайными моментами поступления информации и запросов, случайными моментами возникновения отказов элементов, ошибками операторов и т. п. Имитационное и стохастическое моделирование предлагают мощный инструментарий для решения задач устойчивости работы сервера, его загруженность, расширение или уменьшение пропускной способности. Совокупность всех типов моделирования позволяет создать цифровой двойник серверной инфраструктуры, даёт возможность тестировать различные сценарии нагрузки без воздействия на production-среду. Если рассматривать веб-сервер, как некоторую систему массового обслуживания, клиентские запросы в которой рассматриваются как поток заявок, чаще всего пуассоновский или простейший. Моделирование, как математическое, так и имитационное включает различные параметры: конфигурацию серверного оборудования, алгоритмы балансировки нагрузки, политики кэширования, механизмы обработки запросов, вероятности состояния системы. При этом целесообразно совместное использование наряду с имитационными моделями математические модели, которые представляют клиентские запросы как случайный процесс на основе теории массового обслуживания и позволяют описать процессы поступления, обслуживания и ожидания заявок в системе, как распределение вероятностей состояний системы, количество занятых каналов, то есть сколько запросов на обработке в данное время, число занятых мест в очереди на обслуживание, можно рассматривать, как задержку обслуживания запроса или отказ в обслуживании, если все каналы и места в очереди заняты.

Имитационная модель работы веб-сервера Apache HTTP Server, как системы массового обслуживания рассмотрена в работе [1], в ней проведено сравнение параметров соответствующей сгенерированной системы и Apache HTTP Server с использованием среды GPSS World.

Работа [2] посвящена анализу высокодинамичных условий с помощью теории массового обслуживания, создана соответствующая математическая модель и сравниваются результаты для различных серверов. Показатели производительности при планировании вместимости интернет-сервера оцениваются с использованием различных моделей очередей, путем сравнения таких параметров, как длина очереди, время отклика и время ожидания для различных каналов связи.

Математическое и компьютерное моделирование является эффективным инструментом совершенствования информационного обеспечения. Особую роль выступают вопросы быстродействия выполнения запросов. В работах [3, 4] с использованием аппарата теории массового обслуживания построены математические модели управления надежностью программного обеспечения и компьютерных сетей специального назначения, используемых в системе обеспечения комплексной безопасности учреждений, как управляемой системы массового обслуживания с динамической дисциплиной обслуживания заявок, а также управление очередью M/G/1 с общим временем повторных попыток, допускающим отказы, при этом клиент, обслуживание которого прервано, может оставаться на сервере в ожидании ремонта или уходить и возвращаться, пока сервер ремонтируется, и обслуживание нового не предполагается.

При различных величинах интенсивностей потоков заявок, необходимо производить блокировку высокоинтенсивных потоков необходимо произвести моделирование вероятностных характеристик блокировки запросов на предоставление доступа к ресурсам беспроводной сети. В работе [5] была рассмотрена зависимость вероятности блокировки запроса в зависимости от интенсивности поступления заявок различных типов. Выяснилось, что вероятность блокировки заявки i-го типа имеет вид экспоненциальной функции. По результатам анализа блокировка запросов происходит предсказуемо с учетом характера поступающего трафика.

В данной работе мы рассматриваем загруженность web-сервера, как случайный процесс, с множеством состояний $S = \{s_0, s_1, \dots, s_{c+K}\}$, где s_i – количество занятых каналов, с вероятностью p_i. Он характеризуется тремя потоками, интенсивности могут меняться в зависимости от количества занятых каналов или занятых мест в очереди каждая из которых либо отправляется на обслуживание, либо в очередь, если все каналы заняты.

Первый поток – поток заявок с интенсивностью $\lambda_i(t)$, второй поток – потоком обслуженных заявок с интенсивностью $\mu_i(t)$, если все каналы заняты, то образуется третий поток: очередь, из очереди заявка поступает на обслуживание или покидает очередь, если время ожидание обслуживания вышло, интенсивность освобождения очереди $\eta_m(t)$ (рис. 1).

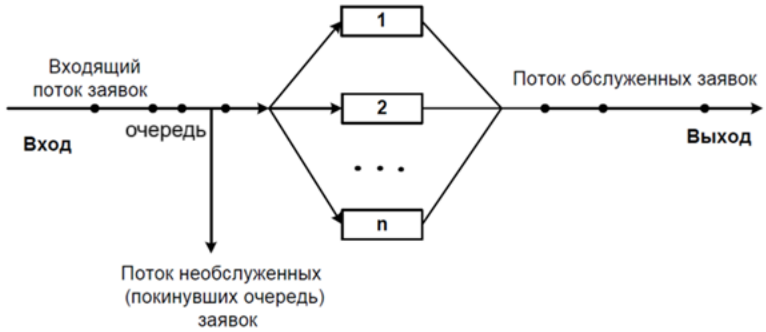

Рис.1 Схема потока обслуживания заявок

Вероятности состояний системы описываются уравнениями Колмогорова [6], при условии нестационарности потоков, в следующем виде, здесь c - количество каналов и K - количество мест в очереди:

- для пустой системы:

$$\frac{dP_0(t)}{dt} = \mu_1(t)P_1(t) - \lambda_0(t)P_0(t);$$

- для системы, в которой занято n каналов:

$$\frac{dP_n(t)}{dt} = \lambda_{n-1}(t)P_{n-1}(t) + \mu_n(t)P_{n+1}(t) - \big(\lambda_n(t) + \mu_n(t)\big)P_n(t);$$

- для системы, в которой заняты все каналы $(n = c)$ и m мест в очереди,

$$\frac{dP_{c+m}(t)}{dt} = \lambda_c(t)P_{c+m-1}(t) - \big(\lambda_c(t) + \mu_c(t) + \eta_m(t)\big)P_{c+m}(t)$$
$$+(\mu_c(t) + \eta_{m+1}(t))P_{c+m+1}(t);$$

- для системы, в которой заняты все каналы и все места в очереди $(n = c, m = K)$:

$$\frac{dP_{c+K}(t)}{dt} = \lambda_c(t)P_{c+K-1}(t) - (\lambda_c(t) + \mu_c(t) + \eta_K(t))P_{c+K}(t).$$

В случае с нашей моделью, количество каналов равно 256, а количество мест в очереди – 511, то есть система может находиться в 768 различных состояниях, следовательно, система уравнений будет состоять из 768 уравнений. Большинство численных методов не подходят для решения систем такой размерности, требуются значительные ресурсы для таких вычислений. Однако, матрица, соответствующая системе уравнений Колмогорова, является трехдиагональной, что позволяет решить систему методом прогонки. Общее число операций в алгоритме прогонки равно $8N + 1$, то есть пропорционально числу уравнений, что делает его экономичным и подходящим для решения систем высокой размерности.

Для решения данной системы дифференциальных уравнений с начальными условиями $P_i(0) = p_i$, $\sum_{i=1}^{c+K} p_i = 1$ можно использовать метод прогонки (метод конечных разностей). Для этого используем неявную схему с шагом τ:

$$\frac{P_0^{k+1} - P_0^k}{\tau} = \mu_1^{k+1} P_1^{k+1} - \lambda_0^{k+1} P_0^{k+1}, \ n = 0$$

$$\frac{P_n^{k+1} - P_n^k}{\tau} = \begin{cases} \lambda_{n-1}^{k+1} P_{n-1}^{k+1} + \mu_n^{k+1} P_{n+1}^{k+1} - (\lambda_n^{k+1} + \mu_n^{k+1}) P_n^{k+1}, & 1 \le n \le c-1, \\ \lambda_c^{k+1} P_{c+m-1}^{k+1} + \mu_{c+m+1}^{k+1} P_{c+m+1}^{k+1} - (\lambda_c^{k+1} + \mu_{c+m}^{k+1}) P_{c+m}^{k+1}, & n = c+m \end{cases}$$

$$\frac{P_{c+K}^{k+1} - P_{c+K}^k}{\tau} = \lambda_c^{k+1} P_{c+k-1}^{k+1} - \mu_{c+k}^{k+1} P_{c+k}^{k+1}.$$

Здесь $\lambda_i^k = \lambda_i(\tau k)$, $\mu_i^k = \mu_i(\tau k), \mu_{c+m}^{k+1} = \mu_c((k+1)\tau) + \eta_m((k+1)\tau)$ – значения интенсивностей потоков входящих и исходящих заявок в момент времени $t_k = \tau k$.

Из аппроксимации получим систему линейных уравнений с трехдиагональной матрицей коэффициентов.

$$\begin{cases} \left(1+\tau\lambda_0^{k+1}\right)P_0^{k+1}-\tau\mu_1^{k+1}P_1^{k+1}=P_0^k, \\ -\tau\lambda_{n-1}^{k+1}P_{n-1}^{k+1}+\left(1+\tau\left(\lambda_n^{k+1}+\mu_n^{k+1}\right)\right)P_n^{k+1}-\tau\mu_n^{k+1}P_{n+1}^{k+1}=P_n^k, 1\le n\le c-1 \\ -\tau\lambda_c^{k+1}P_{c+m-1}^{k+1}+\left(1+\tau\left(\lambda_c^{k+1}+\mu_{c+m}^{k+1}\right)\right)P_{c+m}^{k+1}-\tau\mu_{c+m+1}^{k+1}P_{c+m+1}^{k+1}=P_{c+m}^k, \\ -\tau\lambda_c^{k+1}P_{c+k-1}^{k+1}+\left(1+\tau\mu_{c+k}^{k+1}\right)P_{c+k}^{k+1}=P_{c+k}^k. \end{cases}$$

Метод прогонки состоит из прямого и обратного хода. Вычислим прогоночные коэффициенты α_n и β_n:

$$\begin{cases} \alpha_1=\dfrac{\tau\mu_1^{k+1}}{1+\tau\lambda_0^{k+1}}, \qquad \beta_1=\dfrac{P_0^k}{1+\tau\lambda_0^{k+1}}, \\ \alpha_{n+1}=\dfrac{\tau\mu_n^{k+1}}{1+\tau\left(\lambda_n^{k+1}+\mu_n^{k+1}\right)-\tau\lambda_{n-1}^{k+1}\alpha_n}, \qquad 1\le n\le c-1 \\ \beta_{n+1}=\dfrac{P_n^k+\tau\lambda_{n-1}^{k+1}\beta_n}{1+\tau\left(\lambda_n^{k+1}+\mu_n^{k+1}\right)-\tau\lambda_{n-1}^{k+1}\alpha_n}, \qquad 1\le n\le c-1 \\ \alpha_{c+m+1}=\dfrac{\tau\mu_{c+m+1}^{k+1}}{1+\tau\left(\lambda_c^{k+1}+\mu_{c+m}^{k+1}\right)-\tau\lambda_c^{k+1}\alpha_{c+m}}, \\ \beta_{c+m+1}=\dfrac{P_{c+m}^k+\tau\lambda_c^{k+1}\beta_{c+m}}{1+\tau\left(\lambda_c^{k+1}+\mu_{c+m}^{k+1}\right)-\tau\lambda_c^{k+1}\alpha_{c+m}}, \\ \beta_{c+k+1}=\dfrac{P_{c+k}^k+\tau\lambda_c^{k+1}\beta_{c+k}}{1+\tau\mu_{c+k}^{k+1}-\tau\lambda_c^{k+1}\alpha_{c+k}}. \end{cases}$$

Далее вычисляем вероятности с использованием начального распределения $P_n^0=p_n$, для всех n:

$$P_{c+k}^{k+1}=\beta_{c+k+1}, P_n^{k+1}=\alpha_{n+1}P_{n+1}^{k+1}+\beta_{n+1}, \; n=c+k-1,\dots,0.$$

Для использования в реальных вычислениях нужно будет: задать конкретные функции $\lambda_n(t)$ и $\mu_n(t)$; выбрать шаг по времени (τ) и реализовать алгоритм прогонки на выбранном языке программирования. Моделирование параметров входных и выходных потоков с заданным распределением можно осуществлять методом обратных функций [7] или при помощи встроенных генераторов случайных величин с заданным распределением, например в Python можно использовать модуль numpy.random.

Значения параметров СМО, необходимые, для построения модели, были получены из файлов конфигурации Apache HTTP Server как значения по умолчанию, доступ к которым осуществлялся с помощью программы XAMPP.

Моделирование производительности сервера является важной темой при планировании мощностей и контроле перегрузок веб-серверов. В работе [8] представлена модель очереди веб-сервера Apache, использующую пакетный трафик. Предполагается, что поступление HTTP-запросов представляет собой марковский процесс с модуляцией по Пуассону, а дисциплина обслуживания сервера — это совместное использование ресурсов процессора, если общее количество запросов, которые могут быть обработаны за один раз, ограничено. Получены показатели производительности веб-сервера, такие как среднее время отклика, пропускная способность и вероятность блокировки, с помощью моделирования.

Широкое распространение клиент-серверной технологии взаимодействия и облачных вычислений в настоящий момент времени поднимает вопросы эффективности работы параллельного сервера, а также возможности прогнозировать результаты в зависимости от степени загрузки и характеристик оборудования. В статье [9] производится имитационное моделирование параллельного сервера с отказами в среде AnyLogic, а затем производится многомерная оптимизация методом взвешенной суммы. Для этого построена имитационная модель системы массового обслуживания с отказами, содержащая имитатор работы сервера, терминалы, имитатор отказов и сегменты сбора статистики.

Построение имитационной модели с использованием системы моделирования GPSS World [10], более подробно рассмотрена в работе [1]. Тестирование имитационной модели проведено на основе приложения

Apache JMeter, чаще всего используемое для имитации высокой нагрузки на сервер.

Рассмотрим решение системы дифференциальных уравнений Колмогорова при условии стационарности входных и выходных потоков заявок, пусть интенсивность потока заявок $\lambda = 0{,}01$, интенсивность обслуживания равна интенсивности высвобождения очереди $\mu = \eta = 2{,}5$, то есть моделируем стационарные потоки, получаем систему уравнений (1).

$$\begin{cases} \dfrac{dP_0(t)}{dt} = 2{,}5P_1(t) - 0{,}01P_0(t) \\ \dfrac{dP_1(t)}{dt} = 0{,}01P_0(t) + 2{,}5P_2(t) - 2{,}501P_1(t) \\ \dfrac{dP_2(t)}{dt} = 0{,}01P_1(t) + 2{,}5P_3(t) - 2{,}501P_2(t) \\ \dots \\ \dfrac{dP_{256+0}(t)}{dt} = 0{,}01P_{256+0-1}(t) + 2{,}5P_{256+0+1}(t) - 2{,}501P_{256+0}(t) \\ \dfrac{dP_{256+1}(t)}{dt} = 0{,}01P_{256+1-1}(t) + 2{,}5P_{256+1+1}(t) - 2{,}501P_{256+1}(t) \\ \dots \\ \dfrac{dP_{256+511}(t)}{dt} = 0{,}01P_{256+511-1}(t) - 2{,}5P_{256+511}(t) \end{cases} \quad (1)$$

Рассмотрим реализацию метода прогонки для системы (1). Для аппроксимации производной используем неявную схему (для устойчивости):

$$\frac{P_i^{n+1} - P_i^n}{\tau} = \begin{cases} 2.5P_1^{n+1} - 0.01P_0^{n+1}, i = 0 \\ 0.01P_{i-1}^{n+1} + 2.5P_{i+1}^{n+1} - 2.501P_i^{n+1}, 1 \le i \le 256 + 510, \\ 0.01P_{i-1}^{n+1} - 2.5P_i^{n+1}, \ i = 256 + 511, \end{cases}$$

где τ — шаг по времени, P_i^n — значение P_i на n-м временном слое.

Для каждого временного (n+1)-го слоя получаем систему линейных уравнений:

$$\begin{cases} (1 + 0.01\tau)\mathrm{P}_0^{\mathrm{n}+1} - 2.5\tau\mathrm{P}_1^{\mathrm{n}+1} = \mathrm{P}_0^{\mathrm{n}}, \\ -0.01\tau\mathrm{P}_{i-1}^{n+1} + (1 + 2.501\tau)\mathrm{P}_i^{n+1} - 2.5\tau\mathrm{P}_{i+1}^{n+1} = \mathrm{P}_i^{\tau}, 1 \le \mathrm{i} \le 766, \\ -0.01\tau\mathrm{P}_{766}^{\mathrm{n}+1} + (1 + 2.5\tau)\mathrm{P}_{767}^{\mathrm{n}+1} = \mathrm{P}_{\mathrm{i}}^{\mathrm{n}}. \end{cases}$$

Вычислим прогоночные коэффициенты для решения системы: коэффициенты α_i , β_i(прямой ход):

$$\alpha_1 = \frac{2.5\tau}{1+0.01\tau}, \quad \beta_1 = \frac{P_0^n}{1+0.01\tau},$$

$$\alpha_{i+1} = \frac{2.5\tau}{1+2.501\tau-0.01\tau\alpha_i}, \qquad \beta_{i+1} = \frac{P_i^n+0.01\tau\beta_i}{1+2.501\tau-0.01\tau\alpha_i},$$

$$1 \le i \le 766,$$

$$\beta_{768} = \frac{P_{767}^n+0.01\tau\beta_{767}}{1+2.5\tau-0.01\tau\alpha_{767}}.$$

Вычисляем P_i^{n+1} (обратный ход):

$$P_{767}^{n+1} = \beta_{768}, \; P_i^{n+1} = \alpha_{i+1}P_{i+1}^{n+1}+\beta_{i+1}, \; i = 766,\dots,0.$$

Зададим начальное распределение при условии, что все каналы в начальный момент свободны: $p_0 = 1, p_j = 0, j = 1,\dots,768.$ Используя начальные условия при $t = 0$, $P_0^0 = 1$, $P_i^0 = 0$ для $i \ne 0$ получаем распределение вероятностей состояний системы относительно времени при стационарных входных и выходных потоков. При начальном распределении, когда все каналы свободны, прогоночные коэффициенты $\alpha_i \cong 0.024975$, $\beta_1 \cong 0.999900$, $\beta_2 = 0.000001$, $\beta_i \approx 0$ быстро стабилизируются и формально каналы системы остаются свободными для обработки информации.

При условии, что все каналы заняты, то есть $P_{256} = 1$, остальные равны нулю, вероятности группируются вокруг 256-й и быстро убывают, образуя распределение близкое к нормальному. Система сохраняет концентрацию вероятности вокруг i=256, несмотря на дискретизацию. Коэффициенты: $\alpha_i \approx 0.024975$, постоянны, β_i резко возрастают при $i = 256$, затем убывают, $P_{256}^{n+1} \approx \beta_{257} \approx 0.9756$. Представленная на имитационной модели конфигурация соответствует заявленным параметрам и полностью соответствует математической стохастической модели. Полученные результаты позволяют сделать выводы, что сгенерированные параметры в имитационной модели полностью соответствуют теоретическим расчётам вероятностной модели.

Ни одна заявка не получила отказ, значит можно говорить, что система не переходила в состояние, когда заняты все места в очереди.

Для проверки созданной модели, мы сравнили ее с реальным веб-сервером Apache. С помощью программы XAMPP локально была развернута среда тестирования и запущен Apache HTTP Server.

Имитационная модель адекватно описала работу веб-сервера, несмотря на незначительное различие некоторых характеристик, при этом можно изменять конфигурации настроек Apache, так как у него есть более 700 параметров, и рассмотреть их влияние на функциональность веб-сервера. Для повышения пропускной способности веб-сервера Apache можно подобрать более эффективную конфигурацию, используя методы оптимизации. Математическая модель с нестационарным потоком заявок может быть использована и в высокоинтенсивном трафике, при отношении среднего числа заявок в единицу времени и среднего числа обслуженных заявок близкому к единице.

Литература

1. Попова Т. М., Слободчиков В.А. Моделирование работы web-сервера на основе системы массового обслуживания // Инженерный вестник Дона. 2024. № 8. URL: ivdon.ru/ru/magazine/archive/n8y2024/9438/.

2. Win L. Application of Queuing Theory for Internet Server // Annual University Journal on Innovative Research and Products. 2019. Vol. 2 Issue. 1. pp.508-515.

3. Царькова Е. Г. Математическая модель управления системой массового обслуживания с динамической дисциплиной обслуживания заявок //

Инженерный вестник Дона. 2022. № 5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7638/.

4. Boualem M. Stochastic analysis of a single server unreliable queue with balking and general retrial time // Discrete and Continuous Models and Applied Computational Science. 2020. Vol. 28, No. 4. – pp. 319-326. – DOI: 10.22363/2658-4670-2020-28-4-319-326.

5. Гончаренко С. Н. Моделирование вероятностных характеристик блокировки запросов на предоставление доступа к радиоресурсам беспроводной сети // Инженерный вестник Дона. 2024. № 3. URL: ivdon.ru/ru/magazine/archive/n3y2024/9100/.

6. Петров Ю. В., Аникин С. Н., Юхно С. А. Моделирование случайных величин: учебное пособие. СПб.: Балт. гос. техн. ун-т., 2020. 90 с.

7. Олейникова С. А. Математическое моделирование и системы массового обслуживания. Воронеж: Изд-во ВГТУ, 2021. 90 с.

8. Andersson M, Cao J., Kihl M., Nyberg Ch. Performance Modeling of an Apache Web Server with Bursty Arrival Traffic // Conference: Proceedings of the International Conference on Internet Computing, IC '03, Las Vegas, Nevada, USA, June 23-26, 2003, Vol. 2. URL: researchgate.net/publication/220968222 (дата обращения: 30.06.2025).

9. Сенкевич Л. Б., Сабитов М. А. Имитационное моделирование и оптимизация работы параллельного сервера с отказами в среде AnyLogic // Вестник Тюменского государственного университета. Физико-математическое моделирование. Нефть, газ, энергетика. 2022. Т. 8, № 1. С. 126-143. DOI: 10.21684/2411-7978-2022-8-1-126-143.

10. GPSS World Student Version// URL: gpss-world-student-version.software.informer.com/ (дата обращения 17.06.2025).

References

1. Popova T. M., Slobodchikov V. A. Inzhenernyj vestnik Dona, 2024. № 8. URL: ivdon.ru/ru/magazine/archive/n8y2024/9438/.

2. Win L. Annual University Journal on Innovative Research and Products, 2019. Vol. 2 Issue. 1. pp.508-515.

3. Tsarkova E. G. Inzhenernyj vestnik Dona, 2022. № 5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7638/.

4. Boualem M. Discrete and Continuous Models and Applied Computational Science, 2020. Vol. 28, No. 4. pp. 319-326. DOI: 10.22363/2658-4670-2020-28-4-319-326.

5. Goncharenko S. N., Radimov I. R. Inzhenernyj vestnik Dona. 2024. № 3. URLivdon.ru/ru/magazine/archive/n3y2024/9100/.

6. Petrov Iu.V., Anikin S.N., Iukhno S.A. Modelirovanie sluchainykh velichin: uchebnoe posobie [Modelling of random variables] SPb.: Balt. gos. tekhn. un-t., 2020. 90 p.

7. Olejnikova S. A. Matematicheskoe modelirovanie i sistemy massovogo obsluzhivaniya [Mathematical modeling and queuing systems]. Voronezh: Izd-vo VGTU, 2021. - 90 p.

8. Andersson M, Cao J., Kihl M., Nyberg Ch. Conference: Proceedings of the International Conference on Internet Computing, IC \'03, Las Vegas, Nevada, USA, June 23-26, 2003, Vol. 2. URL: researchgate.net/publication/220968222 (date assessed 30.06.2025).

9. Senkevich L. B., Sabitov M. A. Vestnik Tyumenskogo gosudarstvennogo universiteta. Fiziko-matematicheskoe modelirovanie. Neft', gaz, energetika. 2022. V. 8, № 1. pp. 126-143. DOI: 10.21684/2411-7978-2022-8-1-126-143.

10. GPSS World Student Version. URL: gpss-world-student-version.software.informer.com/ (date assessed 17.06.2025).

Дата поступления: 27.06.2025

Дата публикации: 25.08.2025