

УДК 519.2, 519.872

DOI:10.25729/ESI.2025.40.4.007

Моделирование производительности облачного узла при коррелированном характере нагрузки

Пауль Светлана Владимировна, Назаров Анатолий Андреевич, Лапатин Иван Леонидович, Иванова Алёна Сергеевна

Национальный исследовательский Томский государственный университет,
Россия, Томск, *paulsv82@mail.ru*

Аннотация. В статье предложено исследование математической модели, описывающей функционирование виртуальных машин в облачном узле. Математическая модель представлена в виде системы массового обслуживания с бесконечным числом приборов, каждый из которых соответствует одной виртуальной машине в узле. Моменты начала жизненного цикла каждой машины в узле моделируются марковским модулированным пуассоновским потоком событий, который учитывает коррелированный характер их запуска. Особенностью этой модели является зависимость интенсивности работы каждой виртуальной машины от их суммарного числа в узле. Этот эффект называется «деградацией скорости обслуживания», с помощью которого в модели учитывается конкуренция за ресурсы в облачных узлах. С одной стороны, это позволяет более точно описывать реальные вычислительные системы, с другой стороны, наличие такой зависимости скорости обслуживания от числа виртуальных машин в системе существенно усложняет процесс исследования таких математических моделей. Это приводит к потребности в разработке новых методов анализа систем с учетом деградации скорости обслуживания. Целью предложенной работы является получение аналитических выражений для расчета вероятностных характеристик модели облачного узла в виде системы массового обслуживания с интенсивностью работы виртуальных машин, зависящей от их общего числа в системе.

Ключевые слова: облачный узел, бесконечнолинейная система массового обслуживания, марковский модулированный пуассоновский поток, деградация скорости обслуживания, метод асимптотически диффузионного анализа

Цитирование: Пауль С.В. Моделирование производительности облачного узла при коррелированном характере нагрузки / С.В. Пауль, А.А. Назаров, И.Л. Лапатин, А.С. Иванова // Информационные и математические технологии в науке и управлении, 2025. – № 4(40). – С. 90-101. – DOI:10.25729/ESI.2025.40.4.007.

Введение. Увеличение объемов современных информационных потоков происходит быстрее, чем развитие вычислительной техники. Поэтому все больше клиентов обращаются к облачным хранилищам и центрам обработки данных. Это связано с высокими финансовыми затратами на покупку и обслуживание необходимого количества оборудования требуемой мощности для обработки объема данных, в то время как использование облачных сервисов является экономически более целесообразным. Для компаний-поставщиков облачных услуг остро стоит проблема соблюдения баланса между техническим оснащением работы облачных узлов и сбалансированного распределения запросов от клиентов. Существующие алгоритмы, программы и методы моделирования и прогнозирования нагрузки не всегда решают проблему нехватки ресурсов системы, что приводит к падению производительности и качества обслуживания. При этом возникает так называемый эффект деградации производительности, скорости и качества обслуживания, связанный с неравномерной и недетерминированной нагрузкой, которую дает каждая виртуальная машина (ВМ) на узле. Это оказывает влияние на такие технические показатели, как пропускная способность, время отклика, утилизация ресурсов, надежность, доступность сетей связи. Данный эффект присущ и другим техническим системам с разделяемыми ресурсами, когда физические ресурсы передающего канала в процессе использования разделяются по времени между несколькими информационными потоками. При перегрузке вычислительных узлов, то есть при значительном превышении запрашиваемого объема ресурсов по сравнению с имеющимися,

возникает конкуренция за ресурсы системы, что приводит к ухудшению качества обслуживания отдельно взятых пользователей. Это в конечном итоге может приводить к потере клиентов. С другой стороны, недостаточная загрузка узлов приводит к излишку ресурсов в виде их простоя, что также приводит к финансовым потерям, связанным, например, с покупкой и обслуживанием физических машин, затратами на электроэнергию и др. Таким образом, для эффективного использования системы с разделяемыми ресурсами, к которым относятся современные облачные системы, необходимо обеспечить баланс между загрузкой системы и поддержанием достаточного уровня производительности и качества обслуживания в рамках предоставляемых сервисов.

Для решения этой задачи в предложенной работе с помощью аппарата математического моделирования на основе методов теории массового обслуживания (ТМО) предлагается рассмотреть модель облачного узла в виде бесконечнолинейной системы массового обслуживания с коррелированным входящим потоком (модель Марковского модулированного пуассоновского потока заявок – ММРР) с учетом эффекта возможной деградации производительности и качества обслуживания пользователей. ТМО давно себя зарекомендовала как эффективный инструмент для моделирования поведения и производительности различных технических и экономических систем с разделяемыми ресурсами. Однако, эффект деградации производительности и качества обслуживания пользователей учитывается только в редких работах.

Самый популярный тип моделей, которые учитывают переменные параметры, это системы с параметрами, заданными ступенчатой функцией, когда, например, определяется некое пороговое значение количества ВМ на физической машине. Если количество ВМ меньше порога, то скорость обслуживания имеет одно значение, если больше – другое. Такой подход к моделированию используется в работах [1-3]. Другой метод заключается в использовании моделирования интенсивности обслуживания, интенсивности поступления запросов и времени ожидания (до начала обслуживания) в качестве зависимых случайных величин [4]. Этот метод довольно сложен для понимания и не выглядит полезным, так как на практике не удастся аналитически задать или оценить зависимость между случайными величинами. В работе [5] изучены системы массового обслуживания с произвольным временем обслуживания, продолжительность которого зависит от объема незавершенной работы в системе на момент поступления заявок. Когда в системе имеется большой объем незавершенной работы, вновь прибывшие заявки вынуждены получать обслуживание низкого качества (т. е. короткое время обслуживания). В статье [5] разработан численный метод вычисления распределения вероятностей времени отклика (времени ожидания плюс времени обслуживания). Метод обладает простотой вычисления и широтой области применимости.

В предложенной работе приведено исследование математической модели облачного узла в виде системы массового обслуживания с бесконечным числом приборов, которые моделируют процесс работы ВМ в узле. Предполагается, что все ВМ могут работать на узле одновременно. Производительность каждой ВМ зависит от их общего числа, поскольку узел имеет ограниченные ресурсы (количество ядер ЦП, кеш память различного уровня, оперативную память) для их параллельного выполнения. Этот эффект описывается в системе функцией деградацией, моделирующей процесс снижения интенсивности обслуживания. Применение модели ММРР в качестве входящего потока позволяет учесть несколько режимов запуска ВМ на узле и их коррелированный характер. Событиями входящего потока предлагается моделировать моменты запуска ВМ. Целью исследования является нахождение распределения вероятностей числа ВМ в облачном узле, что позволит вычислять различные вероятностные характеристики предложенной модели.

Математическая модель. Рассмотрим систему массового обслуживания, на вход которой поступает Марковский модулированный пуассоновский поток заявок [6–8], моделирующий моменты запуска ВМ в облачном узле (рисунок 1). Переключение между состояниями в ММРР управляется цепью Маркова с непрерывным временем $m(t) = 0, 1, \dots, M$. Генератор $\mathbf{Q}_{dop} = [q_{vm}]$ для ММРР представляет собой матрицу, которая организована следующим образом: недиагональные элементы положительны и имеют смысл интенсивностей переходов из одного состояния в другое, диагональный элемент – это интенсивности выхода из текущего состояния, взятые с противоположным знаком. Таким образом, длительности пребывания в состояниях ММРР равны длительностям пребывания в состояниях управляющей цепи Маркова и подчиняются экспоненциальному распределению [9–11]. В интервалах, когда ММРР-поток находится в состоянии n , события наступают (запускаются ВМ) с интенсивностью λ_n , которые образуют диагональную матрицу $\mathbf{\Lambda}_{dop} = [\lambda_n]$.

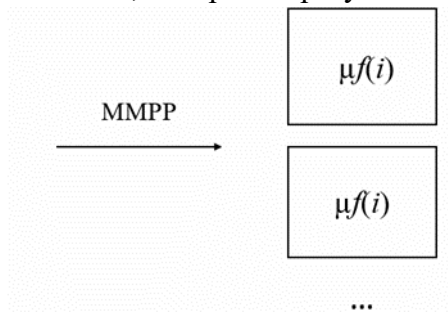


Рис. 1. Бесконечнолинейная система массового обслуживания с ММРР входящим потоком и деградацией скорости обслуживания

Каждая заявка входящего потока мгновенно попадает на прибор и начинает обслуживание, при этом параметр распределения времени обслуживания зависит от числа заявок в системе. Термин «обслуживание на приборе» описывает процесс функционирования в узле одной ВМ. Все ВМ могут работать одновременно на узле. Интенсивность обслуживания (определяющая длительность жизненного цикла ВМ в облачном узле) можно записать, как $\mu f(i)$, где μ – интенсивность работы единственной ВМ в узле (без конкуренции), а $f(i)$ – функция деградации, значения которой являются безразмерным коэффициентом снижения интенсивности работы ВМ в зависимости от их числа i в узле. Чем больше ВМ одновременно находится в облачном узле, тем ниже интенсивность их работы. На практике функцию деградации мы задаем функцией непрерывного аргумента, основываясь на результатах статистического анализа работы реального облачного узла. Например, в качестве функции деградации можно взять убывающую функцию непрерывного аргумента x в виде:

$$\tilde{f}(x) = \frac{1}{\sqrt{1+x}} \text{ (рисунок 2).}$$

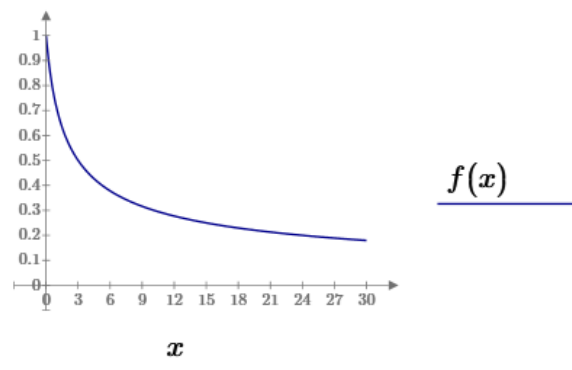


Рис. 2. Пример функции деградации скорости обслуживания

Обозначим процесс $i(t)$ – число заявок в системе (число ВМ в облачном узле) в момент времени t . В предлагаемой работе ставится задача нахождения распределения вероятностей значений процесса $i(t)$. Чтобы записать уравнения, описывающие функционирование облачного узла в терминах теории случайных процессов, необходимо рассматривать марковские процессы, для которых можно составлять системы дифференциальных уравнений Колмогорова. Процесс $i(t)$ не является марковским процессом. Рассмотрим двумерный процесс $\{m(t), i(t)\}$, который уже является марковским процессом, и для его распределения вероятностей:

$$P(m, i, t) = P\{m(t) = m, i(t) = i\} \quad (1)$$

запишем систему дифференциальных уравнений Колмогорова:

$$\begin{aligned} \frac{\partial P(m, i, t)}{\partial t} = & -P(m, i, t)(i\mu f(i) + \lambda_m) + P_m(i-1, t)\lambda_m + \\ & + P(m, i+1, t)(i+1)\mu f(i+1) + \sum_{v=1}^M P(v, i, t)q_{vm} \end{aligned} \quad (2)$$

Решение системы (2) будем искать, применяя метод асимптотически диффузионного анализа в предельном условии высокоинтенсивного ММРР входящего потока, которое обеспечивается параметром T и предельными условиями

$$T \rightarrow \infty \text{ и } Q_{dop} = QT, \Lambda_{dop} = \Lambda T. \quad (3)$$

Предельное условие $T \rightarrow \infty$ и $\Lambda_{dop} = \Lambda T$ высокоинтенсивного входящего потока (высокой интенсивности запуска ВМ в облачном узле) означает, что время между моментами запуска ВМ в узле значительно меньше продолжительности ее жизненного цикла. Таким образом, в узле одновременно работают достаточно большое число ВМ. Условием $Q_{dop} = QT$ соотносятся при $T \rightarrow \infty$ продолжительности интервалов постоянства при запуске ВМ с интенсивностями их запуска в узле и сохраняется структура входящего ММРР, с помощью которого описывается коррелированный характер начала жизненных циклов ВМ в узле.

Метод асимптотически диффузионного анализа реализуется в три этапа. На первом этапе находится функция $a(x)$, которая характеризует среднее число заявок в системе (ВМ в узле). На втором этапе выводим формулы для вычисления функции $b(x)$. На третьем этапе, применяя функции $a(x)$ и $b(x)$, показываем, что предельный процесс, описывающий число заявок в системе (числа ВМ в узле), является диффузионным. Для этого процесса находится предельная плотность распределения, на основе которой выполняется достаточно точная аппроксимация для допредельного дискретного распределения вероятностей числа ВМ в облачном узле с учетом деградации скорости их работы.

Применяя разработанный метод [12, 13], введем следующее обозначение и покажем, как строится функция дискретного аргумента $f(i)$. Будем полагать, что изначально задана функция $\tilde{f}(x)$ непрерывного аргумента x (рисунок 2). Исходную функцию $f(i)$ дискретного аргумента i определим в виде $f(i) = \tilde{f}\left(\frac{i}{T}\right)$. Тогда, система (2) примет вид:

$$\frac{\partial P_m(i, t)}{\partial t} = -P_m(i, t)\left(i\mu\tilde{f}\left(\frac{i}{T}\right) + \lambda_m\right) + P_m(i-1, t)\lambda_m + P_m(i+1, t)(i+1)\mu\tilde{f}\left(\frac{i+1}{T}\right) + \sum_{v=1}^M P_v(i, t)q_{vm}.$$

Последнюю систему перепишем в матричном виде, обозначив вектор-строку $\mathbf{P}(i, t) = \{P(1, i, t), \dots, P(M, i, t)\}$, тогда

$$\frac{\partial \mathbf{P}(i, t)}{\partial t} = \mathbf{P}(i, t)\left(\mathbf{Q}_{dop} - \Lambda_{dop} - i\mu\tilde{f}\left(\frac{i}{T}\right)\mathbf{I}\right) + \mathbf{P}(i-1, t)\Lambda_{dop} + \mathbf{P}(i+1, t)(i+1)\mu\tilde{f}\left(\frac{i+1}{T}\right). \quad (4)$$

Систему (4) перепишем в виде

$$\frac{1}{T} \frac{\partial \mathbf{P}(i, t)}{\partial t} = \mathbf{P}(i, t) \left(\mathbf{Q} - \mathbf{\Lambda} - \frac{i}{T} \mu \tilde{f} \left(\frac{i}{T} \right) \mathbf{I} \right) + \mathbf{P}(i-1, t) \mathbf{\Lambda} + \mathbf{P}(i+1, t) \frac{i+1}{T} \mu \tilde{f} \left(\frac{i+1}{T} \right). \quad (5)$$

Для реализации метода нам понадобится согласованное уравнение, которое получим путем домножения уравнения (5) на единичный вектор столбец \mathbf{e} :

$$\frac{1}{T} \frac{\partial \mathbf{P}(i, t)}{\partial t} \mathbf{e} = -\mathbf{P}(i, t) \left(\mathbf{\Lambda} + \frac{i}{T} \mu \tilde{f} \left(\frac{i}{T} \right) \mathbf{I} \right) \mathbf{e} + \mathbf{P}(i-1, t) \mathbf{\Lambda} \mathbf{e} + \mathbf{P}(i+1, t) \frac{i+1}{T} \mu \tilde{f} \left(\frac{i+1}{T} \right) \mathbf{e}. \quad (6)$$

Решение системы (5)–(6) будем искать, реализуя метод асимптотически диффузионного анализа в предельном условии высокоинтенсивного входящего потока, которое обеспечивается условиями (3).

Первый этап асимптотического анализа. Обозначив $1/T = \varepsilon$, в системе (5)–(6) выполним следующие замены:

$$i/T = \varepsilon i = x, \mathbf{P}(i, t) = \mathbf{P}_1(x, t, \varepsilon). \quad (7)$$

Запишем систему:

$$\varepsilon \frac{\partial \mathbf{P}_1(x, t, \varepsilon)}{\partial t} = \mathbf{P}_1(x, t, \varepsilon) \left(\mathbf{Q} - \mathbf{\Lambda} - x \mu \tilde{f}(x) \mathbf{I} \right) + \mathbf{P}_1(x - \varepsilon, t, \varepsilon) \mathbf{\Lambda} + \mathbf{P}_1(x + \varepsilon, t, \varepsilon) (x + \varepsilon) \mu \tilde{f}(x + \varepsilon), \quad (8)$$

$$\varepsilon \frac{\partial \mathbf{P}_1(x, t, \varepsilon)}{\partial t} \mathbf{e} = -\mathbf{P}_1(x, t, \varepsilon) \left(\mathbf{\Lambda} + x \mu \tilde{f}(x) \mathbf{I} \right) \mathbf{e} + \mathbf{P}_1(x - \varepsilon, t, \varepsilon) \mathbf{\Lambda} \mathbf{e} + \mathbf{P}_1(x + \varepsilon, t, \varepsilon) (x + \varepsilon) \mu \tilde{f}(x + \varepsilon) \mathbf{e}. \quad (9)$$

Решение системы (8)–(9) будем искать в предположении, что существует предел $\lim_{\varepsilon \rightarrow 0} \mathbf{P}_1(x, t, \varepsilon) = \mathbf{P}_1(x, t)$. Устремим параметр $\varepsilon \rightarrow 0$ в матричном уравнении (8), получим однородное матричное уравнение $\mathbf{P}_1(x, t) \mathbf{Q} = 0$, решение которого запишем в виде:

$$\mathbf{P}_1(x, t) = \mathbf{r} \pi_1(x, t), \quad (10)$$

где \mathbf{r} – вектор стационарных вероятностей состояний входящего ММРР, который определяется системой $\mathbf{r} \mathbf{Q}_{dop} = 0$, $\mathbf{r} \mathbf{e} = 1$. Далее вернёмся к уравнению (9), которое запишем с точностью до $O(\varepsilon^2)$, раскладывая в ряд Тейлора по приращению ε аргумента x :

$$\begin{aligned} \varepsilon \frac{\partial \mathbf{P}_1(x, t, \varepsilon)}{\partial t} \mathbf{e} = & -\mathbf{P}_1(x, t, \varepsilon) \left(\mathbf{\Lambda} + x \mu \tilde{f}(x) \mathbf{I} \right) \mathbf{e} + \mathbf{P}_1(x, t, \varepsilon) \mathbf{\Lambda} \mathbf{e} - \varepsilon \frac{\partial}{\partial x} \{ \mathbf{P}_1(x, t, \varepsilon) \mathbf{\Lambda} \mathbf{e} \} + \\ & + \mathbf{P}_1(x, t, \varepsilon) x \mu \tilde{f}(x) \mathbf{e} + \varepsilon \frac{\partial}{\partial x} \{ \mathbf{P}_1(x, t, \varepsilon) x \mu \tilde{f}(x) \mathbf{e} \} + O(\varepsilon^2). \end{aligned}$$

Устремив $\varepsilon \rightarrow 0$ в последнем равенстве, получим:

$$\frac{\partial \mathbf{P}_1(x, t)}{\partial t} \mathbf{e} = -\frac{\partial}{\partial x} \{ \mathbf{P}_1(x, t) (\mathbf{\Lambda} \mathbf{e} - x \mu \tilde{f}(x) \mathbf{e}) \},$$

подставив в которое полученное выше равенство (10), запишем:

$$\frac{\partial \pi_1(x, t)}{\partial t} = -\frac{\partial}{\partial x} \{ \pi_1(x, t) (\mathbf{r} \mathbf{\Lambda} \mathbf{e} - x \mu \tilde{f}(x)) \}, \quad (11)$$

Уравнение (11) является вырожденным уравнением Фоккера–Планка для плотности распределения вероятностей $\pi_1(x, t)$ предельного при $\varepsilon \rightarrow 0$ процесса $\varepsilon i(t)$, который обозначим $x(t)$. Заметим, что процесс $i(t)$ при $\varepsilon \rightarrow 0$ неограниченно возрастает. Процесс $x(t)$ является диффузионным процессом с коэффициентом переноса $a(x)$:

$$a(x) = \mathbf{r} \mathbf{\Lambda} \mathbf{e} - x \mu \tilde{f}(x) \quad (12)$$

и нулевым коэффициентом диффузии. Запишем для него стохастическое дифференциальное уравнение, которое в силу вырожденности уравнения (11) является обыкновенным дифференциальным уравнением

$$dx(t) = (\mathbf{r} \mathbf{\Lambda} \mathbf{e} - x \mu \tilde{f}(x)) dt = a(x) dt. \quad (13)$$

Нас будет интересовать не решение $x(t)$ уравнения (13), а функция $a(x)$. Ниже будет показано, что функция $a(x)$ является коэффициентом переноса вспомогательного (аппроксимирующего) диффузионного процесса для числа ВМ в облачном узле.

Второй этап асимптотического анализа. Реализуя метод асимптотически диффузионного анализа на втором этапе, введем следующее обозначение $1/T = \varepsilon^2$, в системе (5)-(6) выполним замены:

$$i/T = \varepsilon^2 i = x + \varepsilon y, \mathbf{P}(i, t) = \mathbf{P}_2(y, t, \varepsilon). \quad (14)$$

Запишем систему:

$$\begin{aligned} \varepsilon^2 \frac{\partial \mathbf{P}_2(y, t, \varepsilon)}{\partial t} - \varepsilon a(x) \frac{\partial \mathbf{P}_2(y, t, \varepsilon)}{\partial y} = \mathbf{P}_2(y, t, \varepsilon) (\mathbf{Q} - \mathbf{\Lambda} - (x + \varepsilon y) \mu \tilde{f}(x + \varepsilon y) \mathbf{I}) + \\ + \mathbf{P}_2(y - \varepsilon, t, \varepsilon) \mathbf{\Lambda} + \mathbf{P}_2(y + \varepsilon, t, \varepsilon) (x + \varepsilon(y + \varepsilon)) \mu \tilde{f}(x + \varepsilon(y + \varepsilon)), \end{aligned} \quad (15)$$

$$\begin{aligned} \varepsilon^2 \frac{\partial \mathbf{P}_2(y, t, \varepsilon)}{\partial t} \mathbf{e} - \varepsilon a(x) \frac{\partial \mathbf{P}_2(y, t, \varepsilon)}{\partial y} \mathbf{e} = -\mathbf{P}_2(y, t, \varepsilon) (\mathbf{\Lambda} + (x + \varepsilon y) \mu \tilde{f}(x + \varepsilon y) \mathbf{I}) \mathbf{e} + \\ + \mathbf{P}_2(y - \varepsilon, t, \varepsilon) \mathbf{\Lambda} \mathbf{e} + \mathbf{P}_2(y + \varepsilon, t, \varepsilon) (x + \varepsilon(y + \varepsilon)) \mu \tilde{f}(x + \varepsilon(y + \varepsilon)) \mathbf{e}. \end{aligned} \quad (16)$$

Уравнение (15) запишем с точностью до $O(\varepsilon^2)$, раскладывая в ряд Тейлора по приращениям ε аргумента y соответствующие функции:

$$\begin{aligned} -\varepsilon a(x) \frac{\partial \mathbf{P}_2(y, t, \varepsilon)}{\partial y} = \mathbf{P}_2(y, t, \varepsilon) \mathbf{Q} - \varepsilon \frac{\partial}{\partial y} \{ \mathbf{P}_2(y, t, \varepsilon) \mathbf{\Lambda} \} + \\ + \varepsilon \frac{\partial}{\partial y} \{ \mathbf{P}_2(y, t, \varepsilon) (x + \varepsilon y) \mu \tilde{f}(x + \varepsilon y) \} + O(\varepsilon^2). \end{aligned} \quad (17)$$

Решение $\mathbf{P}_2(y, t, \varepsilon)$ системы (17) будем искать в виде:

$$\mathbf{P}_2(y, t, \varepsilon) = \pi_2(y, t) \mathbf{r} + \varepsilon \frac{\partial \pi_2(y, t)}{\partial y} \mathbf{g} + O(\varepsilon^2). \quad (18)$$

Тогда из уравнения (17) получим:

$$\begin{aligned} -\varepsilon a(x) \frac{\partial \pi_2(y, t)}{\partial y} \mathbf{r} = \varepsilon \frac{\partial \pi_2(y, t)}{\partial y} \mathbf{g} \mathbf{Q} - \varepsilon \frac{\partial}{\partial y} \{ \pi_2(y, t) \mathbf{r} \mathbf{\Lambda} \} + \\ + \varepsilon \frac{\partial}{\partial y} \{ \pi_2(y, t) (x + \varepsilon y) \mu \tilde{f}(x + \varepsilon y) \mathbf{r} \} + O(\varepsilon^2). \end{aligned} \quad (19)$$

Перейдем к пределу $\varepsilon \rightarrow 0$ в последнем равенстве, запишем:

$$\frac{\partial}{\partial y} \{ \pi_2(y, t) (\mathbf{g} \mathbf{Q} - \mathbf{r} \mathbf{\Lambda} + x \mu \tilde{f}(x) \mathbf{r} + a(x) \mathbf{r}) \} = 0,$$

откуда получим неоднородную систему линейных алгебраических уравнений:

$$\mathbf{g} \mathbf{Q} = \mathbf{r} (\mathbf{\Lambda} - x \mu \tilde{f}(x) \mathbf{I} - a(x) \mathbf{I}),$$

решение которой запишем в виде разложения:

$$\mathbf{g} = \mathbf{C} \mathbf{r} + \boldsymbol{\varphi},$$

где вектор частных решений $\boldsymbol{\varphi}$ удовлетворяет дополнительному условию $\boldsymbol{\varphi} \mathbf{e} = 0$ и следующему матричному уравнению

$$\boldsymbol{\varphi} \mathbf{Q} = \mathbf{r} (\mathbf{\Lambda} - x \mu \tilde{f}(x) \mathbf{I} - a(x) \mathbf{I}) \quad (20)$$

Вернёмся к уравнению (16), которое запишем с точностью до $O(\varepsilon^3)$, раскладывая по приращениям ε аргумента y соответствующие функции:

$$\begin{aligned} \varepsilon \frac{\partial \mathbf{P}_2(y, t, \varepsilon)}{\partial t} \mathbf{e} - a(x) \frac{\partial \mathbf{P}_2(y, t, \varepsilon)}{\partial y} \mathbf{e} = & - \frac{\partial}{\partial y} \{ \mathbf{P}_2(y, t, \varepsilon) \Lambda \mathbf{e} \} + \frac{\varepsilon}{2} \frac{\partial^2}{\partial y^2} \{ \mathbf{P}_2(y, t, \varepsilon) \Lambda \mathbf{e} \} + \\ & + \frac{\partial}{\partial y} \{ \mathbf{P}_2(y, t, \varepsilon) (x + \varepsilon y) \mu \tilde{f}(x + \varepsilon y) \mathbf{e} \} + \frac{\varepsilon}{2} \frac{\partial^2}{\partial y^2} \{ \mathbf{P}_2(y, t, \varepsilon) (x + \varepsilon y) \mu \tilde{f}(x + \varepsilon y) \mathbf{e} \} + O(\varepsilon^2). \end{aligned}$$

Подставим в последнее равенство разложение (18) и, раскладывая в ряд Тейлора по приращениям εy аргумента x вероятности, с учетом (12) запишем равенство:

$$\begin{aligned} \frac{\partial \pi_2(y, t)}{\partial t} - a(x) \frac{\partial^2}{\partial y^2} \{ \pi_2(y, t) \mathbf{g} \mathbf{e} \} = & - \frac{\partial^2 \pi_2(y, t)}{\partial y^2} [\mathbf{g} (\Lambda - x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e}] + \\ & + \frac{1}{2} \frac{\partial^2 \pi_2(y, t)}{\partial y^2} [\mathbf{r} (\Lambda + x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e}] + \frac{\partial}{\partial y} \left\{ y \pi_2(y, t) \frac{\partial (x \mu \tilde{f}(x))}{\partial x} \right\} + \\ & + \frac{\partial \pi_2(y, t)}{\partial t} = \frac{1}{2} \frac{\partial^2 \pi_2(y, t)}{\partial y^2} [\mathbf{r} (\Lambda + x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e} - 2 \Phi (\Lambda - x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e}] + \\ & + \frac{\partial}{\partial y} \left\{ y \pi_2(y, t) \frac{\partial (x \mu \tilde{f}(x))}{\partial x} \right\}. \end{aligned} \quad (21)$$

Если продифференцировать равенство (12) по x , получим соотношение $a'(x) = - \frac{\partial (x \mu \tilde{f}(x))}{\partial x}$,

которое присутствует в правой части уравнения (21), что позволяет его переписать в более компактном виде:

$$\frac{\partial \pi_2(y, t)}{\partial t} = - \frac{\partial}{\partial y} \{ y \pi_2(y, t) a'(x) \} + \frac{1}{2} \frac{\partial^2 \pi_2(y, t)}{\partial y^2} [\mathbf{r} (\Lambda + x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e} - 2 \Phi (\Lambda - x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e}]. \quad (22)$$

Уравнение (22) является уравнением Фоккера-Планка для плотности распределения вероятностей диффузионного процесса $y(t)$ с коэффициентом переноса $ya'(x)$ и коэффициентом диффузии:

$$b(x) = \mathbf{r} (\Lambda + x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e} - 2 \Phi (\Lambda - x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e}. \quad (23)$$

Тогда для диффузионного процесса $y(t)$ мы можем записать определяющее его стохастическое дифференциальное уравнение:

$$dy(t) = ya'(x)dt + \mathbf{r} \sqrt{(\Lambda + x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e} - 2 \Phi (\Lambda - x \mu \tilde{f}(x) \mathbf{I}) \mathbf{e}} dw(t). \quad (24)$$

тогда уравнение (24) перепишем в более компактном виде:

$$dy(t) = ya'(x)dt + \mathbf{r} \sqrt{b(x)} dw(t). \quad (25)$$

Третий этап асимптотического анализа. На первом и втором этапах реализации метода были сделаны замены (7) и (14), что аналогично рассмотрению некоторого вспомогательного диффузионного процесса:

$$z(t) = x(t) + \varepsilon y(t), \text{ где } \varepsilon = \sqrt{\frac{1}{T}}, \quad (26)$$

на основе которого будет построена диффузионная аппроксимация распределения вероятностей значений процесса $i(t)$, с помощью которого моделируется число ВМ в узле. Для построения аппроксимации сформулируем лемму и теорему.

Лемма. С точностью до $O(\varepsilon^2)$ случайный процесс $z(t)$ является решением стохастического дифференциального уравнения

$$dz(t) = a(z)dt + \sqrt{\varepsilon^2 b(z)}dw(t). \quad (27)$$

Доказательство. Так как диффузионный процесс $x(t)$ является решением стохастического дифференциального уравнения (13), а диффузионный процесс $y(t)$ удовлетворяет уравнению (25), то в силу обозначений (23), (26), для процесса $z(t)$ запишем равенство:

$$dz(t) = d(x(t) + \varepsilon y(t)) = (a(x) + \varepsilon y a'(x))dt + \varepsilon \sqrt{b(x)}dw(t) + O(\varepsilon^2). \quad (28)$$

Коэффициенты в равенстве (28) с учетом (27) запишем в виде разложений с точностью до $O(\varepsilon^2)$

$$a(x) + \varepsilon y a'(x) = a(x + \varepsilon y) + O(\varepsilon^2) = a(z) + O(\varepsilon^2),$$

$$\varepsilon \sqrt{b(x)} = \varepsilon \sqrt{b(x + \varepsilon y) + O(\varepsilon)} = \varepsilon \sqrt{b(z)} + O(\varepsilon^2) = \sqrt{\varepsilon^2 b(z)} + O(\varepsilon^2),$$

С точностью до $O(\varepsilon^2)$ равенство (28) можно переписать в виде:

$$dz(t) = a(z)dt + \sqrt{\varepsilon^2 b(z)}dw(t),$$

которое совпадает с (27). Лемма доказана.

Так как процесс $z(t)$ является решением стохастического дифференциального уравнения, следовательно, этот процесс является диффузионным процессом с коэффициентом переноса $a(z)$ и коэффициентом диффузии $b(z)$. Будем называть его вспомогательным диффузионным процессом, на основе которого будет построена диффузионная аппроксимация распределения вероятностей числа ВМ в облачном узле.

Будем полагать, что исследуемая система функционирует в стационарном режиме, что позволяет рассмотреть стационарную плотность распределения вероятностей процесса $z(t)$:

$$s(z, t) = s(z) = \frac{\partial P\{z(t) < z\}}{\partial z}.$$

Докажем следующее утверждение.

Теорема. Стационарная плотность распределения вероятностей $s(z)$ случайного процесса $z(t)$ имеет вид:

$$s(z) = \frac{C}{b(z)} \exp \left\{ 2T \int_0^z \frac{a(x)}{b(x)} dx \right\}, \quad (29)$$

где C – нормирующая константа.

Доказательство. Сделаем обратные предельные замены:

$$\varepsilon = \frac{1}{T}.$$

Согласно доказанной Лемме, диффузионный процесс $z(t)$ с коэффициентом переноса $a(z)$ и коэффициентом диффузии $b(z)$ является решением стохастического дифференциального уравнения (27), следовательно, его стационарная плотность распределения вероятностей $s(z)$ является решением уравнения Фоккера-Планка:

$$-\frac{\partial}{\partial z} \{a(z)s(z)\} + \frac{1}{2} \frac{\partial^2}{\partial z^2} \left\{ \frac{1}{T} b(z)s(z) \right\} = 0,$$

которое, в свою очередь, является обыкновенным дифференциальным уравнением второго порядка. Решая это дифференциальное уравнение, учитывая краевое условие $s(\infty) = 0$, получим формулу для вычисления плотности распределения вероятностей $s(z)$ нормированного числа заявок в системе, которая совпадает с (29). Теорема доказана.

Точность диффузионной аппроксимации распределения вероятностей числа заявок. Для реализации полученных предельных формул необходимо построить аппроксимацию дискретного распределения вероятностей $P(i)$ числа $i(t)$ ВМ в облачном узле

с ММРР потоком, который описывает моменты запуска ВМ и деградацию скорости их работы. Для перехода от плотности распределения $s(z)$ вспомогательного диффузионного случайного процесса $z(t)$ к дискретному распределению вероятностей значений $P(i)$ случайного процесса $i(t)$ запишем формулу для вычисления неотрицательной функции $G(i)$ дискретного аргумента i в виде:

$$G(i) = \frac{C}{b\left(\frac{i}{T}\right)} \exp \left\{ 2T \int_0^{\frac{i}{T}} \frac{a(u)}{b(u)} du \right\}. \quad (30)$$

Запишем формулу для расчета дискретного распределения вероятностей числа ВМ в узле, принимая во внимание условие нормировки:

$$P(i) = \frac{G(i)}{\sum_{i=0}^{\infty} G(i)}, \quad (31)$$

которую будем называть диффузионной аппроксимацией распределения вероятностей числа ВМ в облачном узле.

Процесс построения аппроксимации изучаемого распределения описывается следующим алгоритмом.

1. Задаются значения параметров матриц изучаемой системы Λ_{dop} , \mathbf{Q}_{dop} , μ , T , функция $\tilde{f}(x)$ непрерывного аргумента x , из которой получаем значения функции дискретного аргумента $f(i) = \tilde{f}\left(\frac{i}{T}\right)$.
2. Находится стационарное распределение вероятностей \mathbf{r} управляющего ММРР потоком процесса $m(t)$, которое определяется системой уравнений $\mathbf{r}\mathbf{Q}_{dop} = 0$, $\mathbf{r}\mathbf{e} = 1$.
3. Вычисляется коэффициент переноса $a(x)$ вспомогательного диффузионного процесса по формуле (12).
4. Находится вектор Φ , как решение матричного уравнения (20) при условии $\Phi\mathbf{e} = 0$.
5. Вычисляется коэффициент диффузии $b(x)$ вспомогательного диффузионного процесса по формуле (23).
6. Вычисляется функция $G(i)$ по формуле (30).
7. Строится дискретная аппроксимация распределения вероятностей $P(i)$ числа заявок на орбите по формуле (31).

Определим следующие матрицы, характеризующие ММРР входящий поток:

$$\Lambda_{dop} = \begin{bmatrix} 0,5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 6 \end{bmatrix}, \quad \mathbf{Q}_{dop} = \begin{bmatrix} -1 & 0,2 & 0,8 \\ 0,5 & -1,5 & 1 \\ 0,8 & 1,2 & -2 \end{bmatrix}.$$

Согласно формулам из работы [14], интенсивность входящего ММРР потока вычисляется по формуле $\mathbf{r}\Lambda_{dop}\mathbf{e} = 1,92$, что характеризует число ВМ, запускаемых на облачном узле в единицу времени. Зададим величину $\mu = 5$, обратную среднему времени жизненного цикла одной ВМ при отсутствии других ВМ на процессоре; параметр высокой интенсивности входящего потока $T = 10$, функцию деградации скорости обслуживания определим в виде

$\tilde{f}(x) = \frac{1}{\sqrt{1+x}}$ и соответствующую предельную функцию деградации $\tilde{f}\left(\frac{i}{T}\right)$. На рисунке 3 представлена диффузионная аппроксимация распределения вероятностей числа занятых

приборов в системе или, что тоже самое, числа ВМ в облачном узле с учетом предложенной функции деградации.

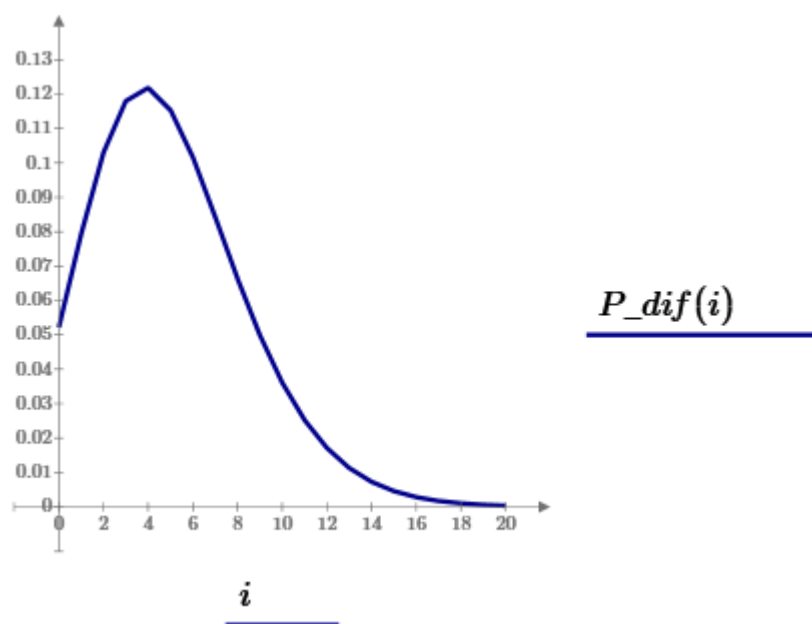


Рис. 3. Диффузионная аппроксимация распределения вероятностей числа ВМ в облачном узле

Как видно из этого примера, при интенсивности запуска ВМ в узле (около 20), распределение вероятностей числа ВМ в облаке имеет большую дисперсию, что приводит к увеличению вероятностей больших значений процесса. Соответственно использование модели входящего потока, которая учитывает не только интенсивность, но и характер запуска ВМ, позволяет более точно получать характеристики моделируемого процесса.

Закключение. В предложенной статье рассмотрена математическая модель облачного узла в виде бесконечнолинейной системы массового обслуживания с учетом зависимости скорости обслуживания заявок от числа занятых приборов, на вход которой на обслуживание поступает марковский модулированный пуассоновский поток сообщений, с помощью которого можно учитывать корреляцию интервалов между моментами запуска ВМ в узле. Ставилась задача нахождения распределения вероятностей числа занятых приборов в системе или, что то же самое, числа ВМ в узле, с помощью которого можно вычислять вероятностные характеристики системы, такие, как среднее и дисперсию, моменты и квантили заданного уровня. Именно число занятых приборов (число ВМ) определяет загрузку вычислительного узла, а соответственно, и уровень снижения скорости обслуживания. Информация о распределении числа занятых приборов напрямую дает нам представление о том, какую долю времени скорость обработки находилась на конкретном уровне. Это позволяет оценивать качество обслуживания при заданных не только интенсивности, но и характере нагрузки за счет использования непуассоновской модели входящего потока. В результате адаптации и реализации метода асимптотически диффузионного анализа было получено предельное в условиях роста числа запущенных ВМ на узле распределение вероятностей числа работающих ВМ в облачном узле и построена его аппроксимация.

Авторы выражают благодарности рецензентам и редакции журнала за изучение статьи и ценные предложения и комментарии.

Список источников

1. Liu X., Li S., Tong W. A Queuing model considering resources sharing for cloud service performance. The Journal of supercomputing, 2015, vol. 71, no. 11, pp. 4042-4055, DOI:10.1007/s11227-015-1503-z.

2. Choudhary A., Chakravarthy S.R., Sharma D.C. Analysis of MAP/PH/1 queueing system with degrading service rate and phase type vacation. *Mathematics*, 2021, vol. 9, no. 19, p. 2387, DOI: 10.3390/math9192387.
3. Goswami V., Patra S., Mund G.B. Performance analysis of cloud with queue-dependent virtual machines. In 2012 1st International Conference on recent advances in information technology (RAIT), IEEE, 2012, pp. 357-362, DOI: 10.1109/RAIT.2012.6194446.
4. Morozov E. A General multiserver state-dependent queueing system. *Centre de Recerca Matematica*, 2010, vol. 927, p. 17.
5. Leung K.K. Load-dependent service queues with application to congestion control in broadband networks. *Performance evaluation*, 2002, vol. 50, no. 1, pp. 27-40, DOI: 10.1016/S0166-5316(02)00045-7.
6. Вишневский В.М. Стохастические системы с коррелированными потоками. Теория и применение в телекоммуникационных сетях / В. М. Вишневский, А.Н. Дудин, В.И. Клименок. – М.: Техносфера, 2018. – С. 564
7. Наумов В.А. О марковских и рациональных потоках случайных событий. I / В.А. Наумов, К.Е. Самуйлов – Информатика и ее применение, 2020. – Т. 14, вып. 3. – С. 13-19.
8. Наумов В.А. О марковских и рациональных потоках случайных событий. II / В.А. Наумов, К.Е. Самуйлов – Информатика и ее применение, 2020. – Т. 14, вып. 4. – С. 37-46.
9. Nazarov A., Moiseev A., Lapatin I. et al. Real waiting time in single-server resource queue with Markovian arrival process. *Communications in computer and information science*, 2023, vol. 1803, pp. 116-125, DOI:10.1007/978-3-031-32990-6_10.
10. Nazarov A., Paul S., Phung-Duc T., et al. Analysis of tandem retrial with common orbit and MMPP incoming flow. *Lecture notes in computer science*, 2022, vol. 13766, pp. 270-283, DOI:10.1007/978-3-031-23207-7_21.
11. Nazarov A., Paul S., Phung-Duc T., et al. Two-way communication retrial queue with markov modulated poisson input and multiple types of outgoing calls. *Communications in computer and information science*, 2022, vol. 1748, pp. 370-381, DOI:10.1007/978-3-031-30648-8_29.
12. Nazarov A., Phung-Duc T., Paul S., et al. Diffusion limit for single-server retrial queues with renewal input and outgoing calls. *Mathematics*, 2022, vol. 10, no. 6, art. no. 948, 14 p., DOI:10.3390/math10060948.
13. Nazarov A., Paul S., Phung-Duc T., et al. Analysis of tandem retrial queue with common orbit and poisson arrival process. *Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics)*, 2021, vol. 13104, pp. 441–456, DOI:10.1007/978-3-030-91825-5_27.
14. Nazarov A., Moiseev A., Lapatin I., et al. Multi-level MMPP as a model of fractal traffic. *Communications in computer and information science*, 2021, vol. 1391, pp. 61-77, DOI:10.1007/978-3-030-72247-0_5.

Пауль Светлана Владимировна. Доктор физико-математических наук, доцент; профессор; каф. теории вероятностей и математической статистики, AuthorID: 533655, SPIN: 6250-1997, ORCID: 0000-0002-3681-0676, paulsv82@mail.ru.

Назаров Анатолий Андреевич. Доктор технических наук, профессор; профессор; каф. теории вероятностей и математической статистики AuthorID: 2954, SPIN: 1341-6014, ORCID: 0000-0002-5097-5629, nazarov.tsu@gmail.com.

Лапатин Иван Леонидович. Кандидат физико-математических наук; доцент; каф. прикладной информатики AuthorID: 587572, SPIN: 1869-3849, ORCID: 0000-0002-1198-2113, ilapatin@mail.ru.

Иванова Алёна Сергеевна. Бакалавр; каф. теории вероятностей и математической статистики. ORCID: 0009-0009-6688-4543, alenkaivas6@gmail.com.

UDC 519.2, 519.872

DOI:10.25729/ESI.2025.40.4.007

Modeling cloud node performance under correlated load conditions

Svetlana V. Paul, Anatoly A. Nazarov, Ivan L. Lapatin, Alyona S. Ivanova

National research Tomsk state university,

Russia, Tomsk, paulsv82@mail.ru

Abstract. The paper proposes a study of a mathematical model describing the operation of virtual machines in a cloud node. The mathematical model is presented as a queueing system with an infinite number of servers, each of which corresponds to a single virtual machine in a node. The lifecycle start times of each machine in a node are modeled by a Markovian Modulated Poisson Process, which takes into account the correlated nature of their launches. A distinctive feature of this model is the dependence of the operational intensity of each virtual machine on their total number in the node. This effect is called "service rate degradation," and it allows the model to account for competition for resources in cloud nodes. While this allows for a more accurate description of real computing

systems, the dependence of service rate on the number of virtual machines in a system significantly complicates the study of such mathematical models. This leads to the need to develop new methods for analyzing systems taking into account the degradation of the service rate and correlations between incoming requests. The purpose of the proposed work is to obtain analytical expressions for calculating the probabilistic characteristics of a cloud node in the form of a queueing system with the rate of processing requests depending on their total number in the system.

Keywords: cloud node, infinite queueing system, Markovian Modulated Poisson Process, service rate degradation, asymptotical diffusion analysis method

References

1. Liu X., Li S., Tong W. A Queueing model considering resources sharing for cloud service performance. The Journal of supercomputing, 2015, vol. 71, no. 11, pp. 4042-4055, DOI:10.1007/s11227-015-1503-z.
2. Choudhary A., Chakravarthy S.R., Sharma D.C. Analysis of MAP/PH/1 queueing system with degrading service rate and phase type vacation. Mathematics, 2021, vol. 9, no. 19, p. 2387, DOI: 10.3390/math9192387.
3. Goswami V., Patra S., Mund G.B. Performance analysis of cloud with queue-dependent virtual machines. In 2012 1st International Conference on recent advances in information technology (RAIT), IEEE, 2012, pp. 357-362, DOI: 10.1109/RAIT.2012.6194446.
4. Morozov E. A General multiserver state-dependent queueing system. Centre de Recerca Matematica, 2010, vol. 927, p. 17.
5. Leung K.K. Load-dependent service queues with application to congestion control in broadband networks. Performance evaluation, 2002, vol. 50, no. 1, pp. 27-40, DOI: 10.1016/S0166-5316(02)00045-7.
6. Vishnevsky V.M., Dudin A.N., Klimenok V.I. Stokhasticheskie sistemy s korrelirovannymi potokami. Teoriya i primeneniye v telekommunikatsionnykh setyakh [Stochastic Systems with Correlated Flows. Theory and Application in Telecommunication Networks]. Moscow, Tekhnosfera Publ., 2018, 564 p.
7. Naumov V.A., Samuylov K.E. O markovskikh i ratsional'nykh potokakh sluchaynykh sobytiy. I [On Markov and Rational Flows of Random Events. I]. Informatika i ee primeneniye [Informatics and its Applications], 2020, vol. 14, iss. 3, pp. 13-19.
8. Naumov V.A., Samuylov K.E. O markovskikh i ratsional'nykh potokakh sluchaynykh sobytiy. II [On Markov and Rational Flows of Random Events. II]. Informatika i ee primeneniye [Informatics and its Applications], 2020, vol. 14, iss. 4, pp. 37-46.
9. Nazarov A., Moiseev A., Lapatin I. et al. Real waiting time in single-server resource queue with Markovian arrival process. Communications in computer and information science, 2023, vol. 1803, pp. 116-125, DOI:10.1007/978-3-031-32990-6_10.
10. Nazarov A., Paul S., Phung-Duc T., et al. Analysis of tandem retrial with common orbit and MMPP incoming flow. Lecture notes in computer science, 2022, vol. 13766, pp. 270-283, DOI:10.1007/978-3-031-23207-7_21.
11. Nazarov A., Paul S., Phung-Duc T., et al. Two-way communication retrial queue with markov modulated poisson input and multiple types of outgoing calls. Communications in computer and information science, 2022, vol. 1748, pp. 370-381, DOI:10.1007/978-3-031-30648-8_29.
12. Nazarov A., Phung-Duc T., Paul S., et al. Diffusion limit for single-server retrial queues with renewal input and outgoing calls. Mathematics, 2022, vol. 10, no. 6, art. no. 948, 14 p., DOI:10.3390/math10060948.
13. Nazarov A., Paul S., Phung-Duc T., et al. Analysis of tandem retrial queue with common orbit and poisson arrival process. Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics), 2021, vol. 13104, pp. 441-456, DOI:10.1007/978-3-030-91825-5_27.
14. Nazarov A., Moiseev A., Lapatin I., et al. Multi-level MMPP as a model of fractal traffic. Communications in computer and information science, 2021, vol. 1391, pp. 61-77, DOI:10.1007/978-3-030-72247-0_5.

Paul Svetlana Vladimirovna. Doctor of Physical and Mathematical Sciences; Associate Professor; Probability Theory and Mathematical Statistics Department. AuthorID: 533655, SPIN: 6250-1997, ORCID: 0000-0002-3681-0676, paulsv82@mail.ru.

Nazarov Anatolij Andreevich. Doctor of Technical Sciences; Professor; Professor; Probability Theory and Mathematical Statistics Department. AuthorID: 2954, SPIN: 1341-6014, ORCID: 0000-0002-5097-5629, nazarov.tsu@gmail.com

Lapatin Ivan Leonidovich. Ph.D. in Physics and Mathematics; Associate Professor; Department of Applied Informatics. AuthorID: 587572, SPIN: 1869-3849, ORCID: 0000-0002-1198-2113, ilapatin@mail.ru

Ivanova Alyona Sergeevna. Bachelor; Probability Theory and Mathematical Statistics Department. ORCID: 0009-0009-6688-4543, alenkaivas6@gmail.com.

Статья поступила в редакцию 30.07.2025; одобрена после рецензирования 01.10.2025; принята к публикации 04.10.2025.

The article was submitted 07/30/2025; approved after reviewing 10/01/2025; accepted for publication 10/04/2025.