УДК 519.248

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ ПРОЦЕССА ПОТРЕБЛЕНИЯ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ ПРИ ПРИМЕНЕНИИ ТЕХНОЛОГИИ ВИРТУАЛЬНЫХ МАШИН

Носова М.Г., Садыков В.Д.

*ФГБОУ ВО «Томский государственный университет систем управления и радиоэлектроники»,
Томск, e-mail: nosovamgm@gmail.com*

Разработана и исследована математическая модель процесса потребления вычислительных ресурсов при выполнении пользовательских задач при условии применения технологии виртуальных машин с целью повышения эффективности использования вычислительных ресурсов. Математическая модель представлена в виде бесконечно линейной системы массового обслуживания с простейшим входящим потоком заявок и экспоненциальным временем обслуживания их на приборах. Процесс функционирования такой системы определяется двумя случайными процессами: числом обслуживаемых запросов и объемом свободных ресурсов. Исследование этих двух процессов выполнено методами теории массового обслуживания. Получена система дифференциальных уравнений, определяющая распределение вероятностей числа обслуживаемых и поступивших запросов. Найден аналитический вид характеристической функции объема свободных ресурсов. Методом моментов найдены основные вероятностные характеристики величины объема свободных ресурсов, а именно математическое ожидание и дисперсия. Предложенная математическая модель процесса потребления вычислительных ресурсов при выполнении пользовательских задач в виде бесконечно линейной системы массового обслуживания с простейшим входящим потоком заявок и экспоненциальным временем обслуживания их на приборах позволяет оценивать процесс изменения объема свободных ресурсов виртуальной машины во времени и анализировать параметры ее производительности.

Ключевые слова: система массового обслуживания, математическое моделирование, виртуальный сервер, оптимизация, виртуальная машина, потребление памяти

DEVELOPMENT AND RESEARCH OF A MATHEMATICAL MODEL FOR THE PROCESS OF CONSUMPTION OF COMPUTER RESOURCES WHEN APPLYING THE VIRTUAL MACHINE TECHNOLOGY

Nosova M.G., Sadykov V.D.

Tomsk State University of Control Systems and Radioelectronics, Tomsk, e-mail: nosovamgm@gmail.com

A mathematical model of the process of consuming computing resources while performing user tasks provided that virtual machine technology is used to increase the efficiency of using computing resources has been developed and investigated. The mathematical model is presented in the form of an infinitely linear queuing system with a simple incoming flow of applications and exponential service time for them on the devices. The process of functioning of such a system is determined by two random processes: the number of discussed requests and the number of free resources. The study of these two processes was carried out using the methods of queuing theory. A system of differential equations is obtained that determines the probability distribution of the number of serviced and received requests. The analytical form of the characteristic function of the volume of free resources is found. Using the method of moments, the main probabilistic characteristics of the number of free resources, namely the mathematical expectation and variance, are found. The proposed mathematical model of the process of consuming computing resources when performing user tasks in the form of an infinitely linear queuing system with the simplest incoming flow of applications and their exponential service time on devices allows us to evaluate the process of changing the amount of free resources of a virtual machine in time and analyze its performance parameters.

Keywords: queuing system, mathematical modeling, virtual server, optimization, virtual machines, memory consumption

Научно-технический прогресс влечет широкое внедрение информационных технологий во все области жизнедеятельности общества. Современные информационные системы осуществляют не только обработку данных, но поддержку различных сервисов и приложений, предоставляющих доступ конечным пользователям. Одной из проблем современных высоконагруженных информационных систем является потребность в больших объемах вычислительных ресурсов. Для обеспечения непрерывного доступа к информационным системам требуется решить проблему эффективного распределения между сервисами ограниченно-го объема ресурсов [1].

На сегодняшний день существует множество предприятий с корпоративной сетью, состоящей из большего числа конечных рабочих мест пользователей и серверного парка. Для рабочих мест пользователей выделяются компьютеры, которые чаще всего уступают по техническим характеристикам компьютерам, используемым в парке серверов. За счет использования множества приложений и размещения данных, в том числе не относящихся к работе, пользователи рабочих мест полностью загружают процессор и занимают все дисковое пространство компьютера. Для удовлетворения постоянно меняющихся бизнес-требований организации должны инвестировать боль-

ше времени и бюджета для масштабирования ИТ-инфраструктуры. Однако достижение этой цели за счет собственных средств и инвестиций не только не является экономически эффективным, но и организации не смогут обеспечить оптимальное использование ресурсов. Эти проблемы вынудили компании искать некоторые новые альтернативные технологические решения. Одной из таких современных технологий являются облачные вычисления на виртуальных машинах, которые направлены на увеличение вычислительной мощности и позволяют функционировать нескольким операционным системам вместе со всеми ее службами на одном физическом компьютере.

Виртуальная машина – виртуальная среда, которая функционирует как виртуальная компьютерная система с собственным процессором, памятью, сетевым интерфейсом и хранилищем, созданная на физическом аппаратном комплексе (расположенном за пределами или в локальной среде). Технология виртуализации позволяет совместно использовать систему со многими виртуальными средами. Гипервизор управляет оборудованием и отделяет физические ресурсы от виртуальных сред. Ресурсы секционируются по мере необходимости от физической среды до виртуальных машин. Когда виртуальная машина запущена, и пользователь или программа направляет запрос, требующий дополнительных ресурсов из физической среды, гипервизор адресует запрос к ресурсам физической системы, чтобы операционная система и приложения виртуальной машины могли получить доступ к общему пулу физических ресурсов. Благодаря применению технологии виртуальных машин серверный парк может быть существенно оптимизирован, что в итоге приведет к уменьшению объема используемых ресурсов и соответственно затрат на их поддержку.

В связи с этим целью данной работы является разработка и исследование математической модели процесса потребления основных вычислительных ресурсов при применении технологии виртуальных машин с целью повышения эффективности их использования.

Модели и методы теории массового обслуживания широко применяются при решении задач о распределении ресурсов [2–6]. В [2] разработана и анализирована облачная архитектура в виде двух систем массового обслуживания, показано, что такие модели могут быть очень полезными для настройки производительности. В [3] представлена модель для анализа производительности, подходящая для исследования качества услуг облачного сервиса, с исполь-

зованием взаимодействующих стохастических моделей. Автор рассмотрел влияние различных параметров, включая скорость прибытия, время обслуживания запросов и его размер. В [4] представлена модель в виде системы массового обслуживания с высокоинтенсивным входящим потоком запросов. Такая модель более реалистична, так как учитывает свойства поступающих запросов и характеристики облачного центра. Авторы описали новое аналитическое приближение для оценки производительности центра облачных вычислений. В [5] представлен обзор моделей облачных вычислений с использованием систем массового обслуживания с различными механизмами распределения ресурсов. Несмотря на обширный обзор моделей в [5], эти модели не являются исчерпывающими, они мотивируют и открывают исследователям еще больше идей для моделирования с использованием систем массового обслуживания.

В работе предлагается новая модель для исследования процесса потребления вычислительных ресурсов при выполнении пользовательских задач виртуальными машинами.

Математическое описание модели

В качестве математической модели процесса потребления виртуальной машиной вычислительных ресурсов при обработке запросов пользователей рассмотрим бесконечно линейную систему массового обслуживания $M|M|∞$ [7]. На вход системы $M|M|∞$ поступает простейший поток запросов интенсивности $λ$, при этом время обслуживания запросов пользователей распределено по экспоненциальному закону с параметром $μ$ (рисунок).

Система массового обслуживания $M/M/∞$

При поступлении запроса пользователя (заявки) на виртуальную машину (система) для его обслуживания выделяется некоторый объем вычислительных ресурсов (например, оперативной памяти), которые полностью освобождаются, когда запрос покидает систему. Если в момент поступления запроса недостаточно ресурсов у виртуальной машины для его обработки, запрос пользователя теряется.

Процесс функционирования системы можно описать двумя случайными процессами: число обслуживаемых запросов $k(t)$ и объем свободных ресурсов $S(t)$. Будем считать, что потенциальный поток запросов пользователей к виртуальной машине является бесконечным.

Рассмотрим процесс изменения состояний случайных процессов $k(t)$ и $S(t)$ с течением времени t . Изменение числа обслуживаемых запросов и объема свободных вычислительных ресурсов происходит в следующих случаях:

1) на виртуальную машину поступает новый запрос. Вероятность того, что за время Δt на виртуальную машину поступит новый запрос, определяется по формуле:

$$\lambda \Delta t + o(\Delta t). \quad (1)$$

Очевидно, что каждый новый запрос связан с захватом некоторого объема ресурсов ϕ для его обработки. Размер ϕ является случайной величиной с функцией распределения $F_\phi(x)$ и моментами $M\{\phi\} = a_1$ и $M\{\phi^2\} = a_2$;

2) в случайный момент времени запрос полностью обрабатывается виртуальной машиной.

Продолжительность обработки запроса определяется функцией распределения:

$$F(x) = 1 - e^{-\mu x}. \quad (2)$$

После обработки запрос с интенсивностью μ покидает виртуальную машину независимо от продолжительности обслуживания других запросов. Тогда за время Δt запрос покидает машину с вероятностью:

$$\mu \Delta t + o(\Delta t). \quad (3)$$

Каждый обслуженный запрос освобождает ресурсы некоторого объема η . Размер η является случайной величиной с функцией распределения F_η и с моментами $M\{\eta\} = c_1$ и $M\{\eta^2\} = c_2$.

Будем предполагать, что при завершении обработки пользовательского запроса вычислительные ресурсы освобождаются полностью, т.е. в том объеме, в котором они были захвачены изначально. Тогда в данных предположениях, очевидно, $a_1 = c_1$, а $a_2 = c_2$.

Определим $i(t)$ как случайный процесс, который характеризует число пришедших

запросов за время t , а $l(t)$ – число запросов, завершающих обслуживание в момент времени t . Отсюда следует, что изменение объема свободных ресурсов можно описать выражением:

$$V(t) = - \sum_{i=1}^{i(t)} \phi + \sum_{l=1}^{l(t)} \eta. \quad (4)$$

Тогда общий объем ресурсов сервера составит:

$$\tilde{V}(t) = V(0) + V(t),$$

где $V(0)$ – начальный объем свободных вычислительных ресурсов.

В дальнейшем, применяя методы теории массового обслуживания [7], а также подход, подробно описанный в [8, 9], проведено исследование двумерного процесса $\{l(t)i(t)\}$. Данный подход включает следующие этапы.

Этап 1. Составляется система дифференциальных уравнений Колмогорова для вероятностей состояний. При составлении системы используется следующее правило. В левой части уравнения стоит производная вероятности i -го состояния. В правой части – сумма произведений вероятностей всех состояний на интенсивности соответствующих потоков событий, минус суммарная интенсивность всех потоков, выводящих систему из данного состояния, умноженная на вероятность данного (i -го состояния). Уравнения Колмогорова дают возможность найти все вероятности состояний как функции времени.

Этап 2. Производится переход от рассмотрения распределения вероятности к его производящей функции. Из системы дифференциальных уравнений Колмогорова записывается дифференциальное уравнение в частных производных первого порядка для производящей функции.

Этап 3. Методом характеристик находится решение дифференциального уравнения, полученного на этапе 2, и записывается аналитический вид производящей функции.

Этап 4. В теории вероятностей и статистике характеристическая функция любой случайной величины полностью определяет ее распределение вероятностей. Характеристическая функция обеспечивает основу альтернативного маршрута для аналитических результатов по сравнению с работой непосредственно с функциями плотности вероятности или функциями распределения. На этом этапе производится переход от рассмотрения производящей функции к рассмотрению характеристической функции. Записывается аналитический вид характеристической функции.

Этап 5. Используя тот факт, что производная k -го порядка в нуле от характеристической функции равна начальному моменту k -го порядка, умноженному на k -ю степень мнимой единицы, находятся моменты первого и вторых порядков (математическое ожидание и дисперсия).

Продemonстрируем ключевые моменты этапов исследования.

Для распределения вероятностей $P(i, l, t) = P\{i(t) = i, l(t) = l\}$ составлена система дифференциальных уравнений Колмогорова:

$$\frac{\partial P(i, l, t)}{\partial t} = -(\lambda + l\mu)P(i, l, t) + \lambda P(i-1, l-1, t) + (l+1)\mu P(i, l+1, t).$$

Откуда для производящей функции вида

$$F(x, y, t) = \sum_{i=0}^{\infty} \sum_{l=0}^{\infty} x^i y^l P(i, l, t)$$

с начальными условиями

$$F(x, y, 0) = \begin{cases} 1, & l=i=0 \\ 0, & \text{в остальных} \end{cases}$$

получено обыкновенное дифференциальное уравнение в частных производных первого порядка:

$$\frac{\partial F(x, y, t)}{\partial t} = -\lambda F(x, y, t) - \mu y \frac{\partial F(x, y, t)}{\partial y} + \lambda xy F(x, y, t) + \frac{\mu \partial F(x, y, t)}{\partial y},$$

решая которое, запишем

$$F(x, y, t) = \exp \left\{ -\frac{\lambda(x-1)(y-1)}{\mu} e^{-\mu t} - \frac{\lambda}{\mu} (y-1) e^{-\mu t} \right\} \times \\ \times \exp \left\{ \frac{\lambda}{\mu} (x-1)(y-1) + \lambda(x-1)t + \frac{\lambda}{\mu} (y-1) \right\}.$$

В результате определен вид производящей функции двумерного распределения $P(i, l, t) = P\{i(t) = i, l(t) = l\}$. Объем свободных вычислительных ресурсов в произвольный момент времени вычисляется по формуле (4). Характеристическая функция этой величины имеет вид:

$$H(u, t) = M(e^{-uV(t)}) = F(\varphi(u), \eta(u), t) = \\ = \exp \left\{ \frac{\lambda}{\mu} \varphi(u)(\eta(u)-1)(1 - e^{-\mu t}) - \lambda(\varphi(u)-1)t \right\},$$

где $\varphi(u) = Me^{ju\varphi_i}$ – характеристическая функция величины φ_i , $\eta(u) = Me^{ju\eta_i}$ – характеристическая функция величины η_i , j – мнимая единица.

Для нахождения вероятностных характеристик математической модели был применен метод моментов, изложенный в [6].

Нахождение вероятностных характеристик числа обслуживаемых запросов

Зная, что производная первого порядка в нуле от характеристической функции равна начальному моменту первого порядка, умноженному на первую степень мнимой единицы, найдем среднее значение свободных ресурсов для виртуальной машины:

$$\left. \frac{1}{j} \frac{\partial H(u, t)}{\partial u} \right|_{u=0} = MV(t).$$

Учитывая, что:

$$\begin{aligned}\varphi(u)|_{u=0} &= 1, & \eta(u)|_{u=0} &= 1, \\ \varphi'(u)|_{u=0} &= -ja_1, & \varphi''(u)|_{u=0} &= j^2a_2, \\ \eta'(u)|_{u=0} &= -jc_1, & \eta''(u)|_{u=0} &= j^2c_2,\end{aligned}$$

проведя несложные преобразования, нашли:

$$MV(t) = \frac{\lambda}{\mu} c_1 (1 - e^{-\mu t}) + \lambda a_1 t,$$

$$M\tilde{V}(t) = V(0) + MV(t),$$

где $V(0)$ – начальный объем свободных ресурсов.

Проведя аналогичные операции для определения второго начального момента, найдем дисперсию:

$$\begin{aligned}\mathcal{D}\tilde{V}(t) &= MV^2(t) - (MV(t))^2 = \frac{1}{j^2} \frac{\partial^2 H(u, t)}{\partial^2 u} \Big|_{u=0} - (MV(t))^2 = \\ &= \frac{\lambda}{\mu} c_2 (1 - e^{-\mu t}) - 2 \frac{\lambda}{\mu} a_1 c_1 (1 - e^{-\mu t}) - \lambda a_2 t.\end{aligned}$$

Учитывая, что вычислительные ресурсы освобождаются полностью, когда запрос покидает систему, т.е. $a_1 = c_1$, $a_2 = c_2$, запишем итоговые выражения:

$$MV(t) = a_1 \left(\frac{\lambda}{\mu} (1 - e^{-\mu t}) + \lambda t \right), \quad (5)$$

$$M\tilde{V}(t) = V(0) + MV(t),$$

где $V(0)$ – начальный объем свободных вычислительных ресурсов,

$$\mathcal{D}\tilde{V}(t) = a_2 \left(\frac{\lambda}{\mu} (1 - e^{-\mu t}) - \lambda t \right) - 2 \frac{\lambda}{\mu} a_1^2 (1 - e^{-\mu t}). \quad (6)$$

Результатом этого раздела являются выражения (5) и (6), определяющие вероятностные характеристики величины объема свободных ресурсов, а именно математическое ожидание и дисперсию.

Заключение

В данной статье была разработана математическая модель процесса потребления ресурсов при обработке пользовательских запросов виртуальной машиной в виде системы массового обслуживания с неограниченным числом обслуживающих приборов, с простейшим входящим потоком заявок и экспоненциальным временем обслуживания их на приборах. Определены вероят-

ностные характеристики величины объема свободных ресурсов сервера, а именно математическое ожидание и дисперсия.

С опорой на полученные в данной работе результаты в дальнейшем планируется усложнить данную модель, тем самым приблизив её максимально к реальным процессам, что создаст базу для дальнейшего изучения работы с ресурсами виртуальной машины с целью повышения эффективности их использования.

Список литературы

1. Гусев О.В., Жуков А.В., Поляков В.В., Поляков С.В. Проблема адекватной оценки производительности веб-серверов в корпоративных сетях на предприятиях ЦБП //

Новые информационные технологии в ЦБП и энергетике: материалы 6-й научно-технической конференции (Петрозаводск, 20–24 сентября 2004 г.). Петрозаводск: АО Metso-Automation, 2004. С. 84–87.

2. Vilaplana J., Solsona F., Teixido I., Mateo J., Abella F., Rius J. A queuing theory model for cloud computing. *The Journal of Supercomputing*. 2014. vol. 69 (1). P. 492–507.

3. Khazaei H., Misic J., Misic V.B. A fine-grained performance model of cloud computing centers. *IEEE Transactions on parallel and distributed systems*. 2012. vol. 24(11). P. 2138–2147.

4. El Aattar M., Hanini M., Oumellal F., Zaaloul A., Haqiq A. Analysis of Queuing System Model for Overflow Tasks Routed to a Cloud Computing Center. *International Journal of Applied Mathematics and Modeling*. 2014. vol. 2. no. 3. P. 28–37.

5. Murugesan R., Elango C., Kannan S. A Status Report on Resource Allocation in Cloud Computing Using Queuing The-

ory. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*. 2014. vol. 3. issue 11. P. 3603–3608.

6. Носова М.Г. Разработка математической модели для оптимизации параметров ОЗУ // Актуальные научные исследования в современном мире. 2019. № 7–1 (51). С. 106–109.

7. Назаров А.А., Терпугов А.Ф. Теория массового обслуживания. Томск: Изд-во НТЛ, 2010. 228 с.

8. Носова М.Г. Математическая модель компании по микрофинансированию со смешанными потоками входящих рисков // Фундаментальные исследования. 2017. № 12–1. С. 207–211.

9. Nosova M.G. Research of a three-phase autonomous queuing system with a Markov Modulated Poisson process. *Information Technologies and Mathematical Modeling (ITMM-2018): Proceedings of 17th International Conference named after A.F. Terpugov (Tomsk, September 10–15, 2018)*. Tomsk: NTL, 2018. P. 33–38.