НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2024 Том 24 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2024 Vol. 24 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-2-249-255
УДК 004. 942

Оценка вероятностно-временных характеристик компьютерной системы с контейнерной виртуализацией

Ван Кю Фунг[1], Владимир Анатольевич Богатырев[2]✉, Николай Сергеевич Кармановский[3], Ван Хиеу Лэ[4]

[1,2,3,4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[2] Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация

[1] phungvanquy97@gmail.com, https://orcid.org/0009-0006-3278-1106
[2] vladimir.bogatyrev@gmail.com✉, https://orcid.org/0000-0003-0213-0223
[3] karmanov50@mail.ru, https://orcid.org/0000-0002-0533-9893
[4] dragon220294@gmail.com, https://orcid.org/0000-0002-9413-5138

Аннотация

Введение. Для компьютерных систем с контейнерной виртуализацией исследована зависимость задержки обслуживания запросов от числа развертываемых контейнеров. Искомая зависимость обусловлена разделением ограниченных вычислительных ресурсов компьютерной системы между активными и неактивными контейнерами, загруженными в системе. **Метод.** В проведенном исследовании предложено комплексное сочетание аналитической модели массового обслуживания, имитационного моделирования и натурных экспериментов. Исследуемая компьютерная система интерпретируется многоканальной системой массового обслуживания с неограниченной очередью. Особенностью предлагаемого подхода является исследование влияния числа сформированных в системе контейнеров на задержки в очереди и интенсивность обслуживания запросов. Каждому контейнеру сопоставляется канал обслуживания, причем для функционирования контейнера в активном и неактивном состояниях требуется использование части общих ресурсов вычислительной системы. При построении модели предполагается, что входной поток простейший, а обслуживание экспоненциальное. Интенсивность обслуживания зависит от числа развернутых контейнеров и от числа запросов в системе. **Основные результаты.** Экспериментально установлена зависимость интенсивности обслуживания от числа активных контейнеров. Исследование выполнено на платформе, основанной на технологии виртуализации Proxmox с фиксированными ресурсами. Для изучения влияния числа активных контейнеров на интенсивность обслуживания в рамках эксперимента развернут однопоточный веб-сервер в виде нескольких контейнеров, управляемый с помощью портативной расширяемой платформы Kubernetes k3s. Результаты расчетов с применением аналитической модели подтверждены результатами имитационного моделирования, реализованного с использованием библиотеки моделирования SimPy на языке программирования Python. На основе проведенных исследований показана необходимость решения задачи оптимизации числа развертываемых в компьютерной системе контейнеров с учетом влияния их числа на задержки обслуживания запросов. **Обсуждение.** Проведенные исследования могут найти применение при проектировании кластерных систем реального времени, критичных к допустимым задержкам ожидания обслуживания запасов, к обеспечению непрерывности вычислительного процесса и к сохранению уникальных данных, накопленных в процессе работы системы. Предложенные подходы могут быть применены при создании отказоустойчивых распределенных компьютерных систем, в том числе функционирующих при накоплении отказов и реконфигурации системы с перераспределением нагрузки (запросов) при динамической миграции и с репликацией контейнеров.

Ключевые слова

система массового обслуживания, контейнер, виртуальная машина, интенсивностью обслуживание, среднее время ожидания, контейнерная виртуализация

Ссылка для цитирования: Фунг В.К., Богатырев В.А., Кармановский Н.С., Лэ В.Х. Оценка вероятностно-временных характеристик компьютерной системы с контейнерной виртуализацией // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 2. С. 249–255. doi: 10.17586/2226-1494-2024-24-2-249-255

Evaluation of probabilistic-temporal characteristics of a computer system with container virtualization

Van Quy Phung[1], Vladimir A. Bogatyrev[2]✉, Nikolay S. Karmanovskiy[3], Van Hieu Le[4]

[1,2,3,4] ITMO University, Saint Petersburg, 197101, Russian Federation

[2] Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation

[1] phungvanquy97@gmail.com, https://orcid.org/0009-0006-3278-1106

[2] vladimir.bogatyrev@gmail.com✉, https://orcid.org/0000-0003-0213-0223

[3] karmanov50@mail.ru, https://orcid.org/0000-0002-0533-9893

[4] dragon220294@gmail.com, https://orcid.org/0000-0002-9413-5138

Abstract

The dependence of request servicing delay on the number of deployed containers is investigated for computer systems with container virtualization. The sought-after dependency is due to the allocation of limited computational resources of the computer system between active and inactive containers loaded in the system. The conducted research proposes a comprehensive combination of analytical queuing model, simulation modeling, and natural experiments. The studied computer system is interpreted as a multi-channel queuing system with an unlimited queue. The peculiarity of the proposed approach is the study of the influence of the number of containers formed in the system on queue delays and request servicing rate. Each container is associated with a service channel, and for the operation of a container in active and inactive states, the use of part of the common resources of the computing system is required. When constructing the model, it is assumed that the input flow is simple, and the service is exponential. The service rate depends on the number of deployed containers and the number of requests in the system. The experimental dependence of service rate on the number of active containers has been established. The experimental study was carried out on a platform based on Proxmox virtualization technology with fixed resources. To study the influence of the number of active containers on service rate within the experiment, a single-threaded web server was deployed in the form of several containers managed using the portable extensible Kubernetes k3s platform. The results of calculations using the analytical model are confirmed by the results of simulation modeling implemented using the SimPy modeling library in the Python programming language. Based on the conducted research, the need to solve the optimization problem of the number of deployable containers in a computer system regarding the influence of this number on request servicing delays is shown. The conducted research can find application in the design of real-time cluster systems critical to acceptable wait service delays, ensuring the continuity of the computational process, and preserving unique data accumulated during the system operation. The proposed approaches can be applied in the creation of fault-tolerant distributed computer systems, including those operating with failure accumulation and system reconfiguration with load (request) redistribution during dynamic container migration and replication.

Keywords

queuing system, container, virtual machine, intensive maintenance, average waiting time, container virtualization

For citation: Phung V.Q., Bogatyrev V.F., Karmanovskiy N.S., Le V.H. Evaluation of probabilistic-temporal characteristics of a computer system with container virtualization. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 2, pp. 249–255 (in Russian). doi: 10.17586/2226-1494-2024-24-2-249-255

Введение

Проектирование отказоустойчивых распределенных компьютерных систем, особенно функционирующих в реальном времени, требует разработки мер по обеспечению надежности и снижению задержек передачи и обработки данных [1–3]. Обеспечение надежности и отказоустойчивости распределенных систем предполагает введение резервирования при передаче, хранении и обработке данных, с консолидацией резервированных ресурсов в кластеры. Современные тенденции в обеспечении высокой производительности, отказоустойчивости и надежности компьютерных систем при требовании и сохранении непрерывности вычислительных процессов, в том числе в реальном времени [4–7], во многом опираются на технологии виртуализации и контейнеризации [1]. Контейнеризация — метод виртуализации на уровне операционной системы, который позволяет упаковать и изолировать приложения в легковесные, мобильные контейнеры. Каждый контейнер содержит все необходимое для запуска приложения, включая код, библиотеки и другие зависимости. Контейнеризация предполагает широкое применение микросервисных архитектур [8], при котором каждая функция или сервис системы реализуется как отдельный компонент, упакованный в контейнер и развертываемый в контейнерной среде. Для управления и развертывания контейнеров могут использоваться платформы Docker [9], Containerd и другие средства.

Для повышения производительности и надежности [10] системы может создаваться множество реплик контейнеров одного и того же сервиса. Однако использование излишнего числа контейнеров может привести к избыточному потреблению ресурсов, а их увеличение выше определенного порога — отрицательно повлиять на производительность и задержки обслуживания в системе. Все это вызывает потребность решения оптимизационной задачи, которое может опираться на сочетание аналитического и имитационного моделирований при натурных экспериментах на объекте.

Моделирование кластерных и облачных систем представляет собой сложную задачу из-за множества аспектов взаимодействия систем обработки, хранения и передачи данных. Математическая модель облачной

системы с контейнерными технологиями представлена в работе [11], а их устойчивость к DDoS-атакам проанализирована в [12]. В работе [13] предложен метод минимизации времени ожидания обслуживания в облачной системе с использованием модели и очереди ограниченной длины. В вышеперечисленных работах использовалась модель многоканальной системы массового обслуживания (СМО) с неограниченной очередью [14]. Отметим, что данная модель не учитывает функциональную зависимость снижения интенсивности обработки запросов из-за разделения ограниченных ресурсов компьютерной системы между контейнерами. Настоящая работа направлена на решение данного упущения путем разработки аналитической модели СМО с переменной интенсивностью обработки запросов, зависящей от общего числа загруженных заданных контейнеров и количества активных контейнеров, задействованных в обслуживании поступающего в компьютерную систему потока запросов. Разработка таких моделей важна для поддержки проектирования кластерных систем реального времени, критичных к допустимым задержкам ожидания обслуживания запасов, в том числе при накоплении отказов и реконфигурации системы с перераспределением нагрузки (запросов), миграцией и репликацией контейнеров, а также уникальных данных, накопленных в процессе функционирования системы.

Модель компьютерной системы с учетом влияния числа загруженных контейнеров

Рассмотрим компьютерную систему с контейнерной виртуализацией. Исследуемая система считается многоканальной СМО неограниченной очереди [15].

Особенностью предлагаемого подхода является исследование влияния на интенсивность обслуживания запросов чисел, сформированных в системе контейнеров, каждый из которых в активном или неактивном состоянии требует часть общих ресурсов вычислительной системы.

Каждому из n каналов обслуживания сопоставим один контейнер. Предположим, что входной поток простейший, а обслуживание экспоненциальное.

Диаграмма состояний и переходов исследуемой системы представлена на рис. 1, на котором состояния при наличии в системе k запросов обозначены как S_k. Очередь неограниченна, поэтому число запросов в системе может быть меньше или больше числа контейнеров. При интенсивности поступления заявок λ, если число запросов в системе k не больше числа каналов (контейнеров), то интенсивность обслуживания будет $k\mu(n, k)$ (при этом очередь пуста, а число активных контейнеров m равно числу запросов в системе k, $m = k$). Если число запросов k в системе больше числа каналов n, то образуется очередь, а интенсивность обслуживания в независимости от числа запросов в очереди $r = k - n$ будет равно $n\mu(n, n)$ (при этом число активных контейнеров m равно общему числу контейнеров в системе n, $m = n$).

Следует отметить, при виртуализации производится разделение общих вычислительных ресурсов компьютерной системы между контейнерами. Таким образом, для оценки вероятностно-временных характеристик обслуживания необходимо установить вид зависимости $\mu(n, m)$, которая характеризует интенсивности обслуживания запросов одним активным контейнером с учетом ее снижения из-за разделения общих вычислительных ресурсов системы между активными контейнерами. Экспериментальное решение этой задачи будет рассмотрено в следующем разделе.

По представленной диаграмме (рис. 1) состояний и переходов определим основные вероятностно-временные характеристики рассматриваемой системы. В результате получим вероятность отсутствия в системе запросов:

$$p_0 = \left(1 + \sum_{i=1}^{n} \frac{\lambda^i}{i! \prod_{j=1}^{i} \mu(n,j)} + \frac{\lambda^n}{n! \prod_{i=1}^{n} \mu(n,i)} \frac{\lambda}{n\mu(n,n) - \lambda} \right)^{-1}. \quad (1)$$

Формула (1) выведена из условия стационарного состояния:

$$\frac{\lambda}{n\mu(n,n)} < 1.$$

Вероятность нахождения в системе k запросов имеет вид:

$$p_k = \begin{cases} \dfrac{\lambda^k}{k! \prod_{i=1}^{k} \mu(n,i)} p_0, \text{ при } k \le n, \\ \dfrac{\lambda^k}{(n\mu(n,n))^{k-n} n! \prod_{i=1}^{n} \mu(n,i)} p_0, \text{ при } k > n. \end{cases}$$

Рассчитаем среднее количество запросов в системе $L_{\text{сист}}$:

$$L_{\text{сист}} = L_{\text{оч}} + L_{\text{об}}, \quad (2)$$

где $L_{\text{оч}}$ и $L_{\text{об}}$ — среднее число запросов в очереди и на обслуживании.

Вычислим значение $L_{\text{об}}$, которое равно среднему числу занятых контейнеров:

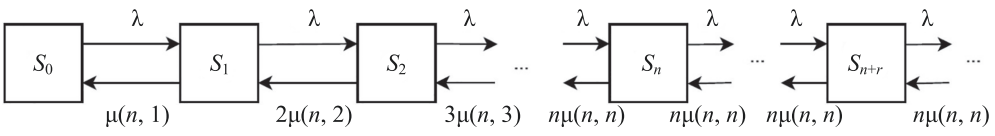

Рис. 1. Граф состояния и переходов исследуемой системы

Fig. 1. State transition graph of the investigated system

$$L_{\text{об}} = \sum_{k=1}^{n} p_k k + \sum_{k=n+1}^{\infty} p_k n = \sum_{k=1}^{n} \frac{k\lambda^k}{k!\prod_{i=1}^{k}\mu(n,i)} p_0 +$$

$$+ n\sum_{r=1}^{\infty} \frac{\lambda^{n+r}}{(n\mu(n,n))^r n!\prod_{i=1}^{n}\mu(n,i)} p_0 = \sum_{k=1}^{n} \frac{k\lambda^k}{k!\prod_{i=1}^{k}\mu(n,i)} p_0 + \quad (3)$$

$$+ \frac{\lambda^{n+1}}{(n-1)!\prod_{i=1}^{n}\mu(n,i)} \frac{1}{n\mu(n,n)-\lambda} p_0.$$

Значение $L_{\text{оч}}$ определим как математическое ожидание количества запросов в очереди r:

$$L_{\text{оч}} = \sum_{r=1}^{\infty} r p_{n+r} = \sum_{r=1}^{\infty} r \frac{\lambda^{n+r}}{(n\mu(n,n))^r n!\prod_{i=1}^{n}\mu(n,i)} p_0 =$$

$$= \frac{\lambda^{n+1}}{n\mu(n,n)n!\prod_{i=1}^{n}\mu(n,i)} p_0 \sum_{r=1}^{\infty} r \frac{\lambda^{r-1}}{(n\mu(n,n))^{r-1}} =$$

$$= \frac{\lambda^{n+1}}{n\mu(n,n)n!\prod_{i=1}^{n}\mu(n,i)} p_0 \sum_{r=1}^{\infty} \frac{d\left(\frac{\lambda}{n\mu(n,n)}\right)^r}{d\left(\frac{\lambda}{n\mu(n,n)}\right)} =$$

$$= \frac{\lambda^{n+1}}{n\mu(n,n)n!\prod_{i=1}^{n}\mu(n,i)} p_0 \frac{d}{d\left(\frac{\lambda}{n\mu(n,n)}\right)} \left(\frac{\frac{\lambda}{n\mu(n,n)}}{1-\frac{\lambda}{n\mu(n,n)}} \right) =$$

$$= \frac{\lambda^{n+1}}{n\mu(n,n)n!\prod_{i=1}^{n}\mu(n,i)} p_0 \frac{(n\mu(n,n))^2}{(n\mu(n,n)-\lambda)^2}.$$

Тогда получим:

$$L_{\text{оч}} = \frac{\lambda^{n+1}}{n\mu(n,n)n!\prod_{i=1}^{n}\mu(n,i)} \left(\frac{n\mu(n,n)}{n\mu(n,n)-\lambda} \right)^2 p_0. \quad (4)$$

Из (2)–(4) по формуле Литтла определим среднее время ожидания в системе:

$$T_{\text{сист}} = \frac{L_{\text{сист}}}{\lambda} = \frac{L_{\text{об}}}{\lambda} + \frac{L_{\text{оч}}}{\lambda},$$

$$T_{\text{сист}} = \sum_{k=1}^{n} \frac{k\lambda^{k-1}}{k!\prod_{i=1}^{k}\mu(n,i)} p_0 +$$

$$+ \frac{\lambda^n}{(n-1)!\prod_{i=1}^{n}\mu(n,i)} \frac{1}{n\mu(n,n)-\lambda} + \quad (5)$$

$$+ \frac{\lambda}{n\mu(n,n)n!\prod_{i=1}^{n}\mu(n,i)} \left(\frac{n\mu(n,n)}{n\mu(n,n)-\lambda} \right)^2 p_0.$$

Экспериментальное определение зависимости интенсивности обслуживания от числа активных контейнеров

Определим зависимость функции интенсивности обслуживания одним активным контейнером $\mu(n, m)$ от общего числа загруженных контейнеров n и количества активных контейнеров m. Искомую функциональную зависимость установим экспериментально. Эксперименты реализованы в лабораторной инфраструктуре, включающей следующие конфигурационные элементы:

— аппаратное обеспечение: 4 × Intel(R) Core(TM) i5-4570 @ 3,20 ГГц;
— версия ядра: Linux 6.2.16-3-pve;
— виртуализация и управление серверами: Proxmox;
— конфигурация и управление контейнером: кластер k3s;
— конфигурация виртуального сервера: одно виртуальное ядро, один виртуальный сокет, 4 ГБ оперативной памяти.

Эксперимент выполнен для однопоточного веб-сервера, упакованного в контейнер и развернутого в нескольких репликах на одном узле кластера Kubernetes k3s (рис. 2).

В результате эксперимента (рис. 3) видно, что изменение числа загруженных (n) и активных контейнеров (m) оказало влияние на интенсивность обслуживания запросов.

Для определения зависимости интенсивности обслуживания одного активного контейнера $\mu(n, m)$ от количества имеющихся n и активных m контейнеров можно использовать различные алгоритмы машинного обучения, спроектированные для решения задачи регрессии. В данной работе такая зависимость найдена с использованием алгоритма Gradient Boosting [16]. Набор данных для обучения имеет следующий вид ($[\mathbf{X}_1,\mathbf{X}_2],\mathbf{Y}$), где $\mathbf{Y}$ — вектор значения интенсивности обслуживания, $\mathbf{X}_1$ и $\mathbf{X}_2$ — векторы со значениями числа заданных контей-

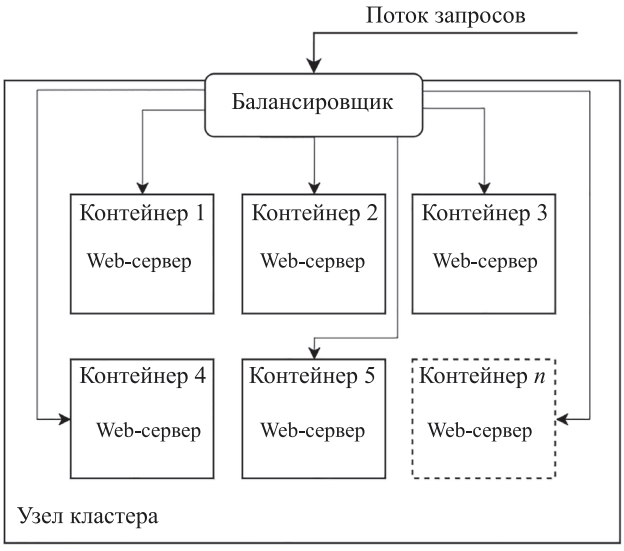

Рис. 2. Схема экспериментальной установки веб-сервера в Kubernetes k3s

Fig. 2. Experimental setup of a web server in Kubernetes k3s

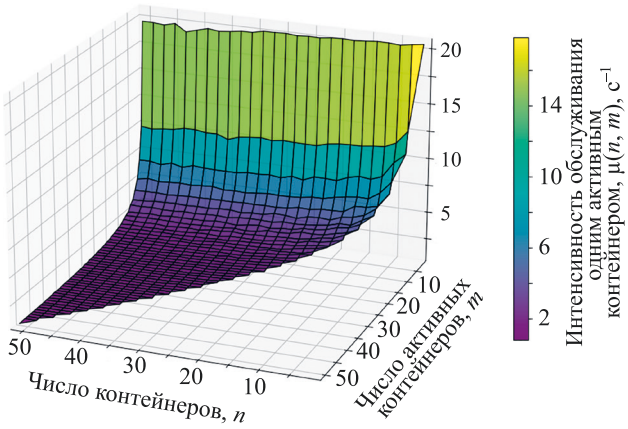

Рис. 3. Результаты эксперимента: корреляция между интенсивностью обслуживания и числом контейнеров

Fig. 3. Experiment results: correlation between service rate and the number of containers

неров и количества активных контейнеров. Алгоритм регрессии выполнен с использованием программы на языке Python с применением библиотеки Sklearn.

График остатков (рис. 4, *b*) отражает разницу между фактическими значениями зависимой переменной и значениями, предсказанными моделью. Красная пунктирная линия, которая проходит через значение «0» оси абсцисс, введена для наглядности представления точности результатов моделирования.

Представленный набор данных случайно разделен на две части: обучающий и тестовый, при этом тестовый набор составляет 20 % от общего объема данных. Результаты регрессии (рис. 4) на тестовом наборе свидетельствуют о высокой точности алгоритма (R-squared: 0,989).

Предел интенсивности запросов для существования стационарного режима обслуживания $\lambda < f(n) = n\mu(n, n)$ получен (рис. 5) из условия $\lambda/n\mu(n, n) < 1$.

Из рис. 5 видно, что при увеличении числа контейнеров способность системы эффективно обрабатывать большие нагрузки возрастает только до определенного предела (примерно 20 контейнеров).

Сравнение результатов аналитического и имитационного моделирований

Для имитационного моделирования исследуемой системы использован инструмент SimPy [17]. SimPy — библиотека для дискретного событийного моделирования, написанная на языке программирования Python. Она предоставляет инструменты для создания и симуляции дискретных событийных моделей, что полезно при исследовании и оптимизации систем, где процессы происходят в дискретные моменты времени. SimPy использует генераторы Python для представления событий и позволяет моделировать процессы, такие как ожидание, обработка задач, переходы состояний и другие дискретные события.

При создании объекта моделирования (многоканальной СМО) в программе симуляции учет для зависимости интенсивности обслуживания от общего

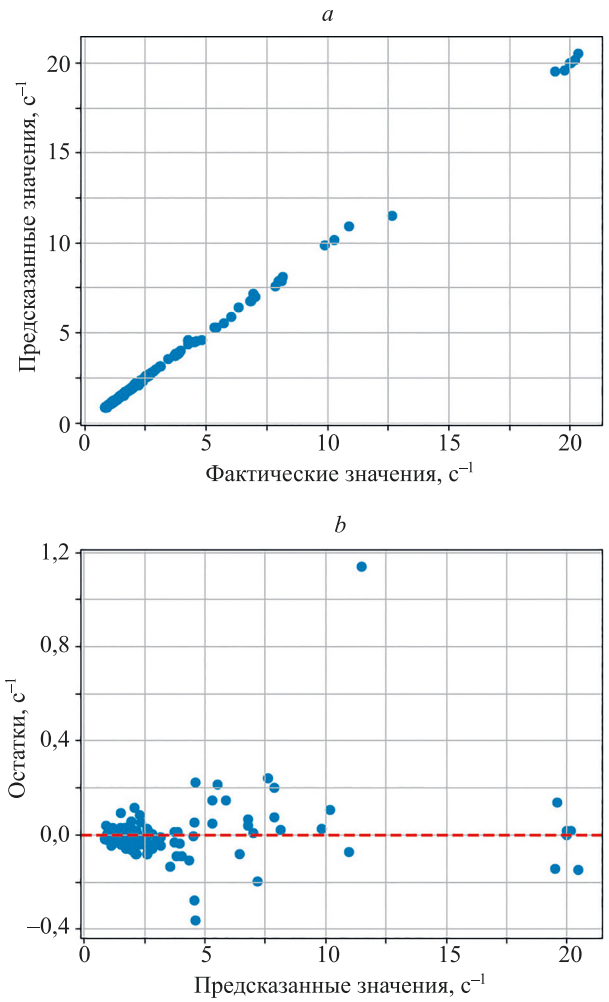

Рис. 4. Результаты тестирования регрессии. Точечные диаграммы фактических и предсказанных значений (*a*); график остатков регрессии (*b*)

Fig. 4. Regression test results: graph of actual and predicted values (*a*); residual plot (*b*)

числа контейнеров и количества активных контейнеров используется функция $\mu(n, m)$, найденная в предыдущем разделе.

На рис. 6 представлен график сравнения результатов симуляции и теоретических расчетов разработанной модели (5) при $\lambda = 15$ с$^{-1}$.

В результате анализа выполненного сравнения данных, представленных на рис. 6, в течение продол-

Рис. 5. Предел интенсивности запросов

Fig. 5. Request rate limit

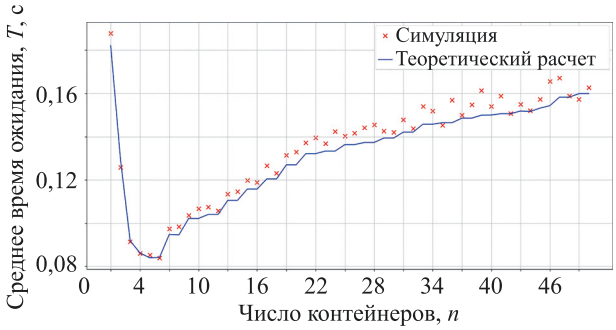

Рис. 6. Сравнение результатов симуляции и теоретического расчета

Fig. 6. Comparison of simulation results and theoretical calculation

жительности временного интервала симуляции, равного 1000 единицам времени, можно отметить, что точность симуляции стремительно приближается к теоретическим предсказаниям (средняя абсолютная ошибка: 2,98 %). В идеальных условиях, при стремлении времени симуляции к бесконечности, точность симуляции практически достигает единицы. Результаты (рис. 6) подтвердили наличие зависимости задержек от числа развертываемых в системе контейнеров, а также существование границы эффективности количества развертываемых машин. До границы при увеличении числа развертываемых контейнеров происходит снижение задержек обслуживания запросов, а после — к их увеличению. Такая закономерность подтверждает необходимость решения задачи оптимизации числа развертываемых контейнеров в зависимости от стоимости и скорости развертывания, а также оценки влияния их числа на надежность энергопотребления компьютерной системы и возникновения задержек обслуживания запросов в ней. Предложенные модели могут быть применены при обосновании построения отказоустойчивых распределенных компьютерных систем, в том числе кластеров, функционирующих в реальном времени при накоплении отказов и реконфигурации системы [18–21].

Заключение

Для компьютерных систем с контейнерной виртуализацией проанализирована зависимость задержки обслуживания запросов вследствие разделения ограниченных ресурсов системы между развернутыми в ней контейнерами. Показано, что по мере увеличения числа загружаемых в систему контейнеров вначале наблюдается снижение задержек обслуживания до некоторой границы, после которой это увеличение приводит к росту задержек обслуживания запросов.

Предложена аналитическая модель компьютерной системы при ее представлении многоканальной системой обслуживания с бесконечной очередью, учитывающей зависимость интенсивности обслуживания от числа представленных и активных контейнеров, разворачиваемых в системе. Зависимость интенсивности обслуживания от количества активных контейнеров установлена экспериментально.

Показана необходимость решения задачи оптимизации числа развертываемых в компьютерной системе контейнеров с учетом влияния этого числа на задержки обслуживания запросов, надежность, энергопотребление, стоимость построения и эксплуатации системы.

Литература

1. Dua R., Raja A.R. Kakadia D. Virtualization vs containerization to support PaaS // Proc. of the 2014 IEEE International Conference on Cloud Engineering. 2014. P. 610–614. https://doi.org/10.1109/IC2E.2014.41
2. Burkov A.A., Rachugin R.O., Turlikov A.M. Analyzing and stabilizing multichannel aloha with the use of the preamble-based exploration phase // Информационно-управляющие системы. 2022. № 5(120). С. 49–59. https://doi.org/10.31799/1684-8853-2022-5-49-59
3. Татарникова Т.М., Архипцев Е.Д. Алгоритм контроллера нечеткой логики для размещения файлов в системе хранения данных // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 6. С. 1171–1177. https://doi.org/10.17586/2226-1494-2023-23-6-1171-1177
4. Астахова Т.Н., Верзун Н.А., Касаткин В.В., Колбанев М.О., Шамин А.А. Исследование моделей связности сенсорных сетей // Информационно-управляющие системы. 2019. № 5(102). С. 38–50. https://doi.org/10.31799/1684-8853-2019-5-38-50
5. Bogatyrev V.A. Increasing the fault tolerance of a multi-trunk channel by means of inter-trunk packet forwarding // Automatic Control and Computer Sciences. 1999. V. 33. N 2. P. 70–76.
6. Татарникова Т.М., Архипцев Е.Д., Кармановский Н.С. Определение размера кластера и числа реплик высоконагруженных информационных систем // Известия высших учебных заведений. Приборостроение. 2023. Т. 66. № 8. С. 646–651. https://doi.org/10.17586/0021-3454-2023-66-8-646-651
7. Bogatyrev V.A. An interval signal method of dynamic interrupt handling with load balancing // Automatic Control and Computer Sciences. 2000. V. 34. N 6. P. 51–57.
8. Hasselbring W., Steinacker G. Microservice architectures for scalability, agility and reliability in E-commerce // Proc. of the 2017

References

1. Dua R., Raja A.R. Kakadia D. Virtualization vs containerization to support PaaS. *Proc. of the 2014 IEEE International Conference on Cloud Engineering*, 2014, pp. 610–614. https://doi.org/10.1109/IC2E.2014.41
2. Burkov A.A., Rachugin R.O., Turlikov A.M. Analyzing and stabilizing multichannel aloha with the use of the preamble-based exploration phase. *Information and Control Systems*, 2022, no. 5(120), pp. 49–59. https://doi.org/10.31799/1684-8853-2022-5-49-59
3. Tatarnikova M.T., Arkhiptsev E.D. Fuzzy logic controller algorithm for placing files in a data storage system. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 6, pp. 1171–1177. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-6-1171-1177
4. Astakhova T.N., Verzun N.A., Kasatkin V.V., Kolbanev M.O., Shamin A.A. Sensor network connectivity models. *Information and Control Systems*, 2019, no. 5, pp. 38–50. (in Russian). https://doi.org/10.31799/1684-8853-2019-5-38-50
5. Bogatyrev V.A. Increasing the fault tolerance of a multi-trunk channel by means of inter-trunk packet forwarding. *Automatic Control and Computer Sciences*, 1999, vol. 33, no. 2, pp. 70–76.
6. Tatarnikova T.M., Arkhiptsev E.D., Karmanovskiy N.S. Determining the cluster size and the number of replicas of highly loaded information systems. *Journal of Instrument Engineering*, 2023, vol. 66, no. 8, pp. 646–651. (in Russian). https://doi.org/10.17586/0021-3454-2023-66-8-646-651
7. Bogatyrev V.A. An interval signal method of dynamic interrupt handling with load balancing. *Automatic Control and Computer Sciences*, 2000, vol. 34, no. 6, pp. 51–57.
8. Hasselbring W., Steinacker G. Microservice architectures for scalability, agility and reliability in E-commerce. *Proc. of the 2017*

IEEE International Conference on Software Architecture Workshops (ICSAW). 2017. P. 243–246. https://doi.org/10.1109/ICSAW.2017.11
9. Hardikar S., Ahirwar P., Rajan S. Containerization: Cloud computing based inspiration technology for adoption through docker and kubernetes // Proc. of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC). 2021. P. 1996–2003. https://doi.org/10.1109/ICESC51422.2021.9532917
10. Богатырев В.А., Богатырев С.В., Богатырев А.В. Оценка готовности компьютерной системы к своевременному обслуживанию запросов при его совмещении с информационным восстановлением памяти после отказов // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23. № 3. С. 608–617. https://doi.org/10.17586/2226-1494-2023-23-3-608-617
11. Srivastava A., Kumar N. Queueing model based dynamic scalability for containerized cloud // International Journal of Advanced Computer Science and Applications (IJACSA). 2023. V. 14. N 1. P. 465–472.
12. Li Z., Jin H., Zou D., Yuan B. Exploring new opportunities to defeat low-rate DDoS attack in container-based cloud environment // IEEE Transactions on Parallel and Distributed Systems. 2020. V. 31. N 3. P. 695–706. https://doi.org/10.1109/TPDS.2019.2942591
13. Pal S., Pattnaik P.K. A simulation-based approach to optimize the execution time and minimization of average waiting time using queuing model in cloud computing environment // International Journal of Electrical and Computer Engineering (IJECE). 2016. V. 6. N 2. P. 743–750. https://doi.org/10.11591/ijece.v6i2.9060
14. Клейнрок Л. Теория массового обслуживания: учебное пособие. М.: Машиностроение, 1979. 432 с.
15. Marshall A.W., Olkin I. A multivariate exponential distribution // Journal of the American Statistical Association. 1967. V. 62. N 317. P. 30–44.
16. Friedman J.H. Greedy function approximation: a gradient boosting machine // The Annals of Statistics. 2001. V. 29. N 5. P. 1189–1132. https://doi.org/10.1214/aos/1013203451
17. Matloff N. Introduction to Discrete-Event Simulation and the SimPy Language. February 13, 2008. 33 p.
18. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Efficiency of servicing heterogeneous traffic when allocating cluster nodes for redundant execution of latency-critical requests // CEUR Workshop Proceedings. 2021. V. 3057. P. 266–273.
19. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Control of multipath transmissions in the nodes of switching segments of reserved paths // Proc. of the 2022 International Conference on Information, Control, and Communication Technologies (ICCT). 2022. P. 1–5. https://doi.org/10.1109/icct56057.2022.9976839
20. Tatarnikova T.M., Sikarev I.A., Bogdanov P.Yu., Timochkina T.V. Botnet attack detection approach in IoT networks // Automatic Control and Computer Sciences. 2022. V. 56. N 8. P. 838–846. https://doi.org/10.3103/s0146411622080259
21. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V. Multipath transmission of heterogeneous traffic in acceptable delays with packet replication and destruction of expired replicas in the nodes that make up the path // Communications in Computer and Information Science. 2023. V. 1748. P. 104–121. https://doi.org/10.1007/978-3-031-30648-8_9

IEEE International Conference on Software Architecture Workshops (ICSAW), 2017, pp. 243–246. https://doi.org/10.1109/ICSAW.2017.11
9. Hardikar S., Ahirwar P., Rajan S. Containerization: Cloud computing based inspiration technology for adoption through docker and kubernetes. *Proc. of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2021, pp. 1996–2003. https://doi.org/10.1109/ICESC51422.2021.9532917
10. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Assessment of the readiness of a computer system for timely servicing of requests when combined with information recovery of memory after failures. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 3, pp. 608–617. (in Russian). https://doi.org/10.17586/2226-1494-2023-23-3-608-617
11. Srivastava A., Kumar N. Queueing model based dynamic scalability for containerized cloud. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2023, vol. 14, no. 1, pp. 465–472.
12. Li Z., Jin H., Zou D., Yuan B. Exploring new opportunities to defeat low-rate DDoS attack in container-based cloud environment. *IEEE Transactions on Parallel and Distributed Systems*, 2020, vol. 31, no. 3, pp. 695–706. https://doi.org/10.1109/TPDS.2019.2942591
13. Pal S., Pattnaik P.K. A simulation-based approach to optimize the execution time and minimization of average waiting time using queuing model in cloud computing environment. *International Journal of Electrical and Computer Engineering (IJECE)*, 2016, vol. 6, no. 2, pp. 743–750. https://doi.org/10.11591/ijece.v6i2.9060
14. Kleinrock L. *Queueing Systems. Vol. 1. Theory*. Wiley, 1974, 448 p.
15. Marshall A.W., Olkin I. A multivariate exponential distribution. *Journal of the American Statistical Association*, 1967, vol. 62, no. 317, pp. 30–44.
16. Friedman J.H. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 2001, vol. 29, no. 5, pp. 1189–1132. https://doi.org/10.1214/aos/1013203451
17. Matloff N. *Introduction to Discrete-Event Simulation and the SimPy Language*. February 13, 2008. 33 p.
18. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Efficiency of servicing heterogeneous traffic when allocating cluster nodes for redundant execution of latency-critical requests. *CEUR Workshop Proceedings*, 2021, vol. 3057, pp. 266–273.
19. Bogatyrev V.A., Bogatyrev S.V., Bogatyrev A.V. Control of multipath transmissions in the nodes of switching segments of reserved paths. *Proc. of the 2022 International Conference on Information, Control, and Communication Technologies (ICCT)*, 2022, pp. 1–5. https://doi.org/10.1109/icct56057.2022.9976839
20. Tatarnikova T.M., Sikarev I.A., Bogdanov P.Yu., Timochkina T.V. Botnet attack detection approach in IoT networks. *Automatic Control and Computer Sciences*, 2022, vol. 56, no. 8, pp. 838–846. https://doi.org/10.3103/s0146411622080259
21. Bogatyrev V.A., Bogatyrev A.V., Bogatyrev S.V. Multipath transmission of heterogeneous traffic in acceptable delays with packet replication and destruction of expired replicas in the nodes that make up the path. *Communications in Computer and Information Science*, 2023, vol. 1748, pp. 104–121. https://doi.org/10.1007/978-3-031-30648-8_9

Авторы

Фунг Ван Кю — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0006-3278-1106, phungvanquy97@gmail.com

Богатырев Владимир Анатольевич — доктор технических наук, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; профессор, Санкт-Петербургский государственный университет аэрокосмического приборостроения, Санкт-Петербург, 190000, Российская Федерация, SC 7006571069, https://orcid.org/0000-0003-0213-0223, vladimir.bogatyrev@gmail.com

Кармановский Николай Сергеевич — кандидат технических наук, доцент, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57192385103, https://orcid.org/0000-0002-0533-9893, karmanov50@mail.ru

Лэ Ван Хиеу — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0002-9413-5138, dragon220294@gmail.com

Authors

Van Quy Phung — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0006-3278-1106, phungvanquy97@gmail.com

Vladimir A. Bogatyrev — D.Sc., Professor, ITMO University, Saint Petersburg, 197101, Russian Federation; Professor, Saint Petersburg State University of Aerospace Instrumentation, Saint Petersburg, 190000, Russian Federation, SC 7006571069, https://orcid.org/0000-0003-0213-0223, vladimir.bogatyrev@gmail.com

Nikolay S. Karmanovskiy — PhD, Associate Professor, Associate Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57192385103, https://orcid.org/0000-0002-0533-9893, karmanov50@mail.ru

Van Hieu Le — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0002-9413-5138, dragon220294@gmail.com

Статья поступила в редакцию 02.02.2024
Одобрена после рецензирования 15.02.2024
Принята к печати 17.03.2024

Received 02.02.2024
Approved after reviewing 15.02.2024
Accepted 17.03.2024