Автоматизация и моделирование в проектировании и управлении. 2026. № 1 (31). С. 58-67.
ISSN 2658-3488 print, ISSN 2658-6436 online
Automation and modeling in design and management. 2026. № 1 (31). P. 58-67.

Научная статья
Статья в открытом доступе
УДК 519.85
doi: 10.30987/2658-6436-2026-1-58-67

ОПТИМИЗАЦИЯ И ЭФФЕКТИВНОСТЬ КОМПЬЮТЕРНОЙ СИСТЕМЫ С КОНТЕЙНЕРНОЙ ВИРТУАЛИЗАЦИЕЙ В УСЛОВИЯХ НЕОПРЕДЕЛЕННОСТИ

Александра Анатольевна Чудинова
Университет ИТМО, г. Санкт-Петербург, Россия
https://orcid.org/0000-0003-4171-6964

Аннотация. *Разработана формула оптимизации для вычисления интенсивности обслуживания и количества активных контейнеров на основе аксиоматического анализа теории массового обслуживания. Цель исследования – проанализировать вероятность интенсивности, рассчитать интенсивность в условиях неопределенности трафика, оптимизировать распределение запросов и выявить связь между оптимизацией и эффективностью. Современные распределённые компьютерные системы, использующие контейнерную виртуализацию, сталкиваются с проблемами производительности и стоимости в условиях неопределенного трафика запросов. Динамичность скорости поступления запросов и взаимозависимость контейнеров, использующих общие вычислительные ресурсы, приводят к трудностям в поддержании стабильности системы, низком времени отклика и эффективном использовании ресурсов. Существующие модели часто упрощают эти взаимосвязи или игнорируют колебания трафика. Поэтому необходима формальная модель, позволяющая в режиме реального времени рассчитывать и корректировать интенсивность обслуживания и количество контейнеров, основанная на теории очередей и подходящая для неопределённых условий трафика. Разработана новая формула на основе теории массового обслуживания, позволяющая через вычислительно-программную трансформацию, рассчитать интенсивность в определенный промежуток времени и определить оптимизацию и эффективность существующей распределительной компьютерной системы с контейнерной виртуализацией. Результаты демонстрируют возможность распределения запросов эффективно с учетом требуемых ресурсов и возможностью анализа поведения распределительной компьютерной системы в определенный промежуток времени. Экспериментальное исследование в режиме реального времени с наблюдением в определенное время суток трафика запросов к Apache JMeter, который системно представляет собой кластер с контейнерной виртуализацией, демонстрирует стабильное распределение запросов балансировщиком и применимость математической модели для проведения вычислений оптимизации и эффективности распределенной компьютерной системы.*

Ключевые слова: теория массового обслуживания, условия неопределенности, оптимизация, эффективность, трафик запросов, распределенные компьютерные системы, *Apache JMeter*, контейнерная виртуализация

Благодарности. Работа выполнена при научном руководстве доктора технических наук, профессора Владимира Анатольевича Богатырева.

Для цитирования: Чудинова А.А. Оптимизация и эффективность компьютерной системы с контейнерной виртуализацией в условиях неопределенности // Автоматизация и моделирование в проектировании и управлении. 2026. №1 (31). С. 58-67. doi: 10.30987/2658-6436-2026-1-58-67.

Original article
Open Access Article

OPTIMIZATION AND EFFICIENCY OF A CONTAINERIZED VIRTUALIZATION COMPUTER SYSTEM UNDER UNCERTAINTY

Alexandra A. Chudinova
ITMO University, Saint Petersburg, Russia
https://orcid.org/0000-0003-4171-6964

Abstract. *The paper has developed an optimization formula to calculate service intensity and the number of active containers based on axiomatic analysis of queuing theory. The aim of the study is to analyse traffic intensity probabilities, calculate intensity under uncertain traffic conditions, optimize query distribution, and establish the relationship between optimization and efficiency. Modern distributed computing systems employing container virtualization encounter performance and cost challenges under uncertain traffic loads. The dynamism of request arrival speeds and interdependence among containers sharing common computational resources lead to difficulties maintaining system stability, low latency, and efficient resource utilization. Existing models often oversimplify these interconnections or*

*ignore traffic fluctuations. Therefore, a formal model is needed to enable real-time calculation and adjustment of service intensity and container counts based on queueing theory, suitable for uncertain traffic conditions. A new formula derived from queuing theory allows computationally transforming service intensity within a specific timeframe and optimizing existing distributed computer systems with container virtualization. The results demonstrate the possibility of effectively distributing queries considering required resources and analysing the behaviour of distributed computer systems within a specific timeframe. Real-time experimental research observing traffic requests via Apache JMeter at a certain time of day, representing a cluster with container virtualization, illustrates stable query distribution by the load balancer and the applicability of the mathematical model for conducting optimization and efficiency computations in distributed computer system*s.

Keywords: queuing theory, uncertainty conditions, optimization, efficiency, query traffic, distributed computer systems, *Apache JMeter*, container virtualization

Acknowledgments. This work was carried out under the scientific supervision of Doctor of Technical Sciences, Professor Vladimir Anatolyevich Bogatyrev.

For citation: Chudinova A.A. Optimization and Efficiency of a Containerized Virtualization Computer System Under Uncertainty. Automation and modeling in design and management, 2026, no. 1 (31). pp. 58-67. doi: 10.30987/2658-6436-2026-1-58-67.

Введение

Теория массового обслуживания является фундаментальным знанием по реализации распределения очередей (потоков) [6]. Задача увеличения скорости обслуживания и распределения ресурсов является актуальной в силу появления новых компьютерных систем, требующих обоснованной дистрибуции потока данных. В случае с кластерными системами требуется учет технических ресурсов и учет экономической рентабельности обслуживания каждого запроса. Для того чтобы система стала производительной и эффективной, требуется учитывать интенсивность трафика запросов, произвести аналитические вычисления возможности оптимизации распределения запросов во избежание простоев в компьютерной системе и предотвратить перегрузку системы в целом и обеспечить возможность быстрой обработки обслуживания запросов. На данный момент экспериментальные исследования сконцентрированы на тестировании системы при различных технических характеристиках компьютерной системы, что несколько ограничивает возможность универсализации математических моделей и требует иных технических реализаций. В теоретическом анализе исследование ставит целью вывода математической формулы оптимизации системы с контейнерной виртуализацией в условиях неопределенности, что делает возможным обеспечить сбалансированное распределение очередей в различных системах.

Материалы, модели, эксперименты и методы

Согласно постановке задачи, интенсивность обслуживания $\mu(n)$ зависит от количества активных контейнеров n, как предполагают уравнения в качестве исходных данных Ку Фунга и В.А. Богатырева [1, 5]. Основываясь на экспериментальных наблюдениях, функция интенсивности обслуживания $\mu(n)$ может быть аппроксимирована функцией (1), такой как:

$$\mu(n) = \mu_0 + \alpha f(n), \quad (1)$$

где μ_0 – базовая интенсивность при отсутствии активных контейнеров, $f(n)$ – функция, моделирующая влияние количества контейнеров, α – экспериментальный коэффициент.

Применимость данной формулы аргументирована базовой системой массового обслуживания с экспоненциальным временем между прибытиями и обслуживаниями и одним сервером, где производительность системы (например, время ожидания $W(n)$ зависит от количества поступающих запросов (2), моделируемых следующим образом [3]:

$$W_n = W_0 + \alpha \cdot f(n), \quad (2)$$

где W_0 – базовое время ожидания, когда нет задач (например, система работает с полной эффективностью или не имеет очереди), α – константа, которая представляет чувствительность системы к количеству поступающих задач, $f(n)$ – функция количества задач (аналогично для количества контейнеров), поступающих в систему, например, $f(n)=n$ (для линейного увеличения количества задач) или $f(n)=\log(n)$ (для убывающей отдачи по мере поступления большего количества задач).

Время ожидания в очереди зависит как от базового значения W_0, так и от возрастающего влияния количества задач на производительность системы, причем влияние модулируется α.

Выбором для $f(n)$ является степенная функция, экспоненциальная функция или функция логистического типа в зависимости от реальных экспериментальных данных.

Рассмотрим систему автомасштабируемых облачных серверов, где количество активных серверов n зависит от входящего трафика. Количество активных контейнеров n следует распределению Пуассона (3) [2]:

$$P(n) = (e^{-\lambda}\lambda^n/n!), \tag{3}$$

при котором:

$$\mu = n \cdot \mu_0, \frac{dn}{d\mu} = 1/\mu_0. \tag{4}$$

Поэтому, если число активных контейнеров n следует распределению вероятностей $P(n)$ (5), то вероятность заданной интенсивности $\mu(n)$ можно записать как [9]:

$$P(\mu) = P(n) \cdot |dn/d\mu|^{-1}, \tag{5}$$

где $P(n)$ получено из эмпирических данных. Неопределенность трафика означает, что λ (интенсивность поступления запросов) является случайной величиной.

Стохастическая модель, такая как процесс Пуассона с параметрами неопределенности, может быть представлена:

$$\lambda = \lambda_0 + \varepsilon(t), \tag{6}$$

где $\varepsilon(t)$ представляет неопределенность нагрузки. Формула (6) расширяет теорию очередей, моделируя неопределенность нагрузки интенсивности поступления запросов, учитывая случайные изменения трафика с течением времени, что влияет на стабильность системы, длину очереди и оптимизацию ресурсов, исходя из стандартной скорости поступления λ в очереди M/M/1 (марковское поступление, марковское обслуживание, один сервер) [8]:

$$P(N(t) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} = n \cdot \mu_0. \tag{7}$$

Ожидаемая интенсивность обслуживания в условиях неопределенности тогда равна:

$$E[\mu] = \int \mu(n) P(n) dn. \tag{8}$$

Формула (8) представляет собой ожидаемое (среднее) значение скорости обслуживания μ, рассчитанное как непрерывное средневзвешенное значение по всем возможным количествам контейнеров n. Здесь $\mu(n)$ – это скорость обслуживания как функция n, а $P(n)$ – вероятность каждого количества контейнеров n.

Оптимизация соотношения интенсивности и эффективности требует определения эффективности (апробирована взаимосвязью между скоростью обслуживания и пропускной способностью запросов, связанных доступной емкостью и обработанными запросами с помощью известных показателей производительности, таких как использование, время отклика и продолжительность очереди) как [10]:

$$\eta = \frac{\text{Доступная мощность}}{\text{Обработанные запросы}}. \tag{9}$$

Следует, что формула (9) неверна, так как переворачивает фундаментальное и общепринятое определение эффективности и ведет себя контринтуитивно при применении к производительности системы в контексте очередей. Ее использование напрямую противоречило бы самой цели оптимизации эффективности, которая заключается в максимизации выхода (10) относительно доступных ресурсов. Поэтому требуется продолжить теоретический анализ и сформировать новую формулу (11).

Оптимизация вычисляется по формуле с учетом таких ограничений, как условия стабильности очереди и максимальные пределы обработки:

$$\max_{\mu(n)} \eta(\mu, n). \tag{10}$$

Эффективность (η) в системе массового обслуживания часто измеряется с точки зрения использования (насколько правильно используются ресурсы системы) и производительности обслуживания (насколько быстро обрабатываются запросы) [4]:

– высокая интенсивность трафика ($\rho \approx 1$) → система почти на полной мощности, что приводит к задержкам и возможным заторам;

– низкая интенсивность трафика ($\rho \ll 1$) → система не продуктивна это означает, что ресурсы (например, серверы, контейнеры) используются неэффективно.

Оптимальная эффективность достигается за счет балансировки скорости обслуживания (μ) и скорости поступления заявок (λ) для минимизации длины очереди (13) и максимального использования ресурсов. Эффективность определяется как:

$$\eta = \frac{\text{Обработанные запросы}}{\text{Доступная мощность}}. \quad (11)$$

Эффективность возрастает по мере приближения к полной загрузке, но если ρ превышает определенный порог ($\rho > 1$), система становится перегруженной и неэффективной:

$$\eta = \rho = \frac{\lambda}{\mu}. \quad (12)$$

Данное соотношение (12) определяет загрузку системы η или ρ как отношение скорости поступления заявок λ к скорости обслуживания μ. Она количественно характеризует степень загрузки системы, показывая, насколько эффективно используются системные ресурсы.

Для оптимизации системы требуется настроить такие параметры, как количество активных контейнеров (n) и скорость обслуживания (μ), чтобы обеспечить $\rho < 1$ (в противном случае очереди будут расти бесконечно). Минимизированная длина очереди и задержки из закона Литтла:

$$L = \lambda W, \quad (13)$$

где L – средняя длина очереди, W – среднее время ожидания.

Стабильно оптимизированная система поддерживает L (13) на низком уровне, сохраняя при этом высокую эффективность. Оптимальное количество серверов/контейнеров (n^*):

$$\frac{d}{dn\left(\frac{\lambda}{n\mu_0}\right)} = 0. \quad (14)$$

Формула (14) используется для нахождения оптимального количества контейнеров n, минимизирующего или стабилизирующего нагрузку на систему, при условии, что λ и μ_0 – константы.

В условиях неопределенности трафика λ колеблется, требуя адаптивной оптимизации. Динамическое распределение ресурсов отрегулируется n на основе спроса в реальном времени. Стохастическая оптимизация прогнозирует λ и оптимизирует $\mu(n)$ для обработки неопределенности нагрузки. Оптимизация на основе вероятности вычисляется по формуле (15), обеспечивая стабильность системы в различных условиях:

$$E(\rho) = \sum_{n=1}^{\infty} \frac{\lambda}{n\mu_0} P_n, \quad (15)$$

где λ – интенсивность поступления запросов, μ_0 – скорость обслуживания одного сервера, n – количество активных серверов (или контейнеров), P_n – вероятность наличия n контейнеров активными в данный момент времени.

Таким образом, эффективность измеряется с использованием интенсивности трафика (ρ), балансировки загрузки и задержки. Оптимизация определяется корректировкой n (контейнеры), μ (интенсивность обслуживания) и λ (интенсивность поступления запросов) для поддержания стабильности и минимизации затрат. Адаптивная стратегия заключается в динамическом распределении ресурсов, которое помогает поддерживать эффективность в условиях неопределенности.

Требуется построить графики соотношений времени ожидания и количества контейнеров, числа контейнеров и предела интенсивности запросов на основании реализованного исследования, но с учетом адаптивной стратегии по эффективности и оптимизации.

Результаты

Техническими требованиями к экспериментальному исследованию является интернет-соединение, так как предусматривается исследование интенсивности запросов на *Apache JMeter* [7].

Для начала требуется определить интенсивность в различное время суток, чтобы определить вероятность интенсивности.

Вероятность интенсивности позволит повысить надежность системы, прогнозируя и формируя распределение ресурсов. Согласно графику на рис. 1, который основан на формуле (4), получается график с тремя линиями, каждая из которых представляет вероятность интенсивности для разного времени суток.

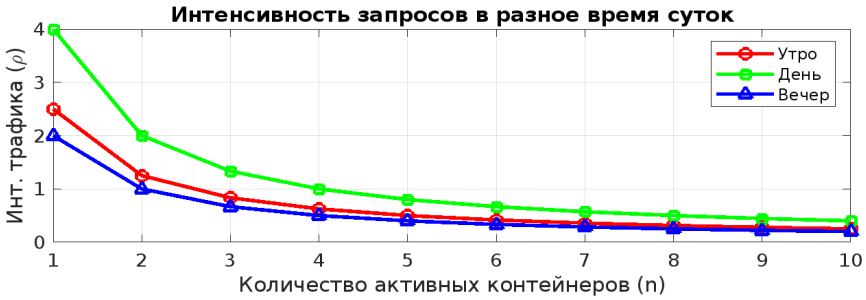

Рис. 1. Интенсивность запросов в разное время суток к Apache JMeter
Fig. 1. Request intensity at different times of day to Apache JMeter

Поскольку в этом случае формула упрощается до постоянного значения, график будет показывать плоские линии при значении 1 для каждого периода времени (утро, день, вечер), но подход может быть расширен или изменен для более сложных сценариев, где $P(\mu)$ не является тривиально 1. Это возможно при обнаружении, как вероятность нахождения системы в состоянии занятости зависит от интенсивности трафика. По мере увеличения ρ (что означает, что система находится в условиях высокой нагрузки) вероятность нахождения в состоянии занятости уменьшается, поскольку система достигает точки насыщения и не может обрабатывать больше запросов:

$$P(\mu) = \frac{1}{(1+\rho)}, \tag{16}$$

где ρ – интенсивность трафика, определяемая как:

$$\rho = \frac{\lambda}{n \cdot \mu_0}, \tag{17}$$

По мере увеличения количества активных контейнеров (n) система становится менее вероятной в состоянии занятости, поскольку большее количество контейнеров позволяет лучше обрабатывать входящие запросы. Вероятность интенсивности $P(\mu)$ уменьшается по мере увеличения интенсивности трафика (16). При более высокой нагрузке трафика (больше λ или меньше n) система с большей вероятностью будет находиться в состоянии загруженности, а вероятность уменьшается.

Таким образом, на рис. 2 представлено два графика для демонстрации поведения системы при различной вероятности интенсивности.

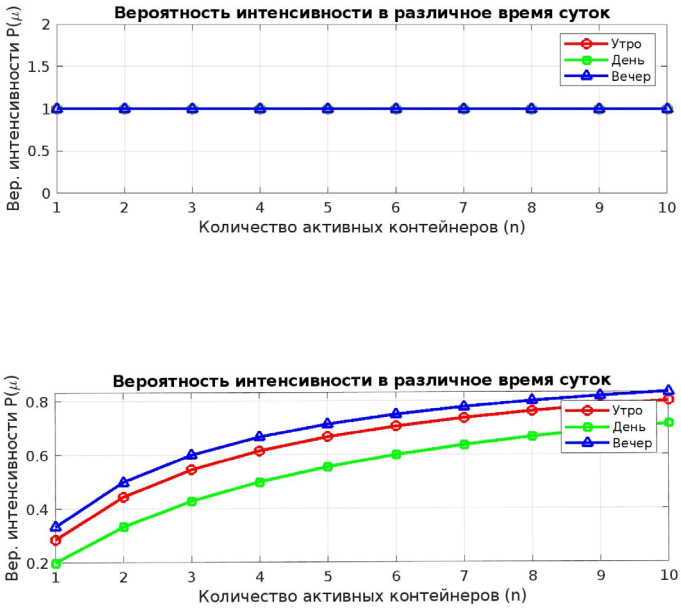

Рис. 2. Вероятность интенсивности запросов в определенный промежуток времени суток
Fig. 2. Probability of query intensity in a certain time period of the day

Верхний график рассчитывается по формуле (15) и представляет собой вероятность того, что сервер будет простаивать, или вероятность пустой очереди в базовой системе массового обслуживания M/M/1, особенно когда ρ < 1. Она связана со стационарной вероятностью P_0 (вероятностью отсутствия клиентов в системе). По мере увеличения интенсивности трафика (ρ) (что означает, что система становится более загруженной) эта вероятность уменьшается, что логично – более загруженная система с меньшей вероятностью будет простаивать. По мере увеличения количества контейнеров (n) интенсивность трафика (ρ) уменьшается (поскольку (17)). Уменьшение ρ приводит к увеличению $P(\mu)$. Это означает, что с увеличением количества контейнеров вероятность того, что система будет менее интенсивной или будет иметь доступную мощность (например, простаивающие серверы), увеличивается.

В нижнем графике формула (16) упрощается до $P(\mu) = 1$ для всех значений ρ (поскольку ρ / ρ = 1 и |1|= 1). Это постоянное значение, равное 1. Здесь (5) предполагает попытку связать функцию плотности вероятности скорости обслуживания с функцией массы вероятности количества контейнеров, получая:

$$P(\mu) = \rho \cdot \left|\frac{1}{\rho}\right|. \quad (18)$$

Однако реализация (18) не выполняет это более сложное преобразование для получения значимого распределения вероятностей. Линии на этом графике плоские на уровне $y = 1$ для любого количества контейнеров и любого времени суток. Это указывает на то, что данный расчет, в текущей реализации, не предоставляет полезной или изменяющейся вероятности интенсивности. Подход нижнего графика предоставляет осмысленную вероятностную метрику, связанную с состоянием системы, в частности, с вероятностью наличия в системе свободной мощности или меньшей загруженности. График помогает понять, как добавление контейнеров влияет на вероятность того, что система сможет немедленно обрабатывать новые запросы без постановки в очередь. Если вы хотите обеспечить определенный уровень свободной мощности (например, для обработки внезапных всплесков нагрузки), наблюдение за этой вероятностью для разных значений n будет очень ценным. Мониторинг производительности может служить индикатором работоспособности системы. Постоянно низкое значение $P(\mu)$ (если интерпретировать его как вероятность простоя) может сигнализировать о надвигающейся перегрузке, что побуждает к действиям по масштабированию. Наблюдая, как эта вероятность изменяется с n, можно найти баланс между избыточным количеством простаивающих ресурсов и недостаточным для эффективного удовлетворения спроса.

Расчет, лежащий в основе нижнего графика, является явно более полезным и теоретически обоснованным для представления вероятности, связанной с интенсивностью системы или ее свободной мощностью в контексте систем массового обслуживания. Он дает представление о том, как добавление контейнеров влияет на вероятность того, что система сможет эффективно обрабатывать запросы. Расчет второго графика, как написано, дает постоянное значение и не предоставляет никаких значимых сведений о динамике системы.

Рис. 3 демонстрирует, как среднее время ожидания входящих запросов уменьшается по мере добавления контейнеров в систему.

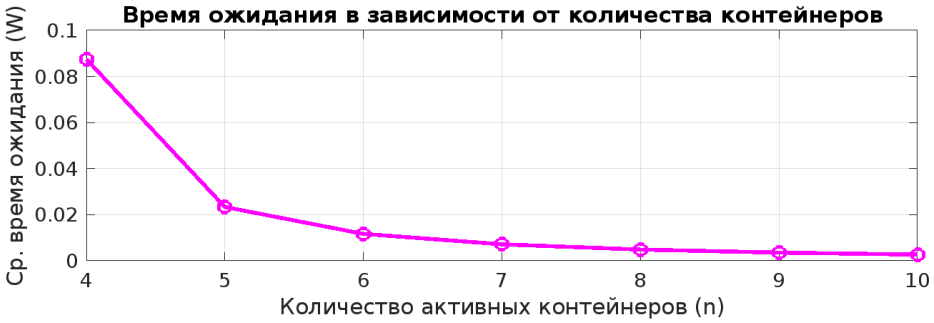

Рис. 3. Время ожидания в зависимости от количества контейнеров

Fig. 3. Waiting time depending on the number of containers

Он показывает, что, хотя увеличение количества контейнеров повышает производительность, после определённого момента этот эффект снижается и для поддержания стабильности системы необходимо поддерживать интенсивность трафика $\rho < 1$.

График на рис. 4 показывает максимальную интенсивность запросов, которую может обработать контейнерная система при увеличении числа активных контейнеров, исходя из условия устойчивости $\rho = 1$.

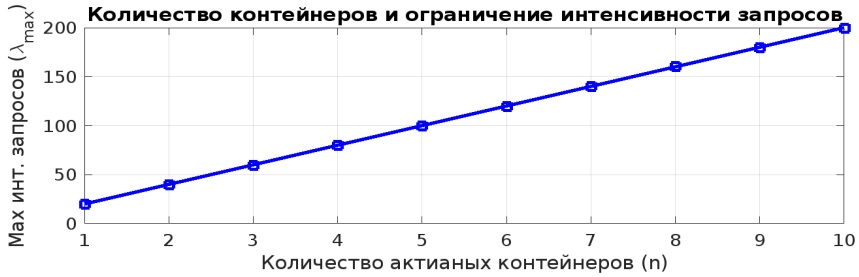

Рис. 4. Количество контейнеров и ограничение интенсивности запросов

Fig. 4. Number of containers and request rate limitation

Он иллюстрирует прямую линейную зависимость, т.е. мощность системы линейно масштабируется с числом контейнеров, что помогает определить, сколько экземпляров необходимо для обработки заданной нагрузки без перегрузки.

Данное исследование показывает, что оптимизация производительности и экономической эффективности распределенной контейнерной системы при переменной нагрузке требует динамического баланса между интенсивностью трафика (ρ), частотой запросов (λ) и количеством активных контейнеров (n). Интенсивность трафика, определяемая формулой (17), должна оставаться строго ниже 1 для обеспечения стабильности системы и предотвращения чрезмерного времени ожидания. Максимальная частота запросов на контейнер устанавливает верхний предел приемлемого входящего трафика, определяя распределение вычислительных ресурсов (19).

$$\lambda_{\max} = n \cdot \mu_0. \tag{19}$$

С помощью аналитического моделирования и графического анализа показано, что стоимость системы, включающая как фиксированные, так и переменные компоненты, может быть минимизирована путем выбора оптимального количества контейнеров. Эта оптимальная точка находится там, где интенсивность трафика близка к насыщению, но не достигает его, что обеспечивает высокую эффективность (12) при одновременном избежании перегрузки и ненужного расхода ресурсов. Время ожидания, обратно пропорциональное разнице $(1-\rho)$, служит чувствительным индикатором снижения производительности, особенно когда ρ приближается к 1. Таким образом, оптимизация системы достигается путем настройки количества контейнеров для поддержания субкритической нагрузки трафика, минимизации времени отклика и снижения эксплуатационных расходов без ущерба для пропускной способности или стабильности.

Согласно рис. 5, эффективность (η) наиболее высока, когда доступно меньше контейнеров. По мере увеличения количества контейнеров (n) эффективность постепенно снижается. График ограничен 1, поскольку эффективность не может превышать 100 % (т.е. $\rho > 1$ приводит к перегрузке системы). Если эффективность слишком низкая, это означает, что ресурсы используются недостаточно.

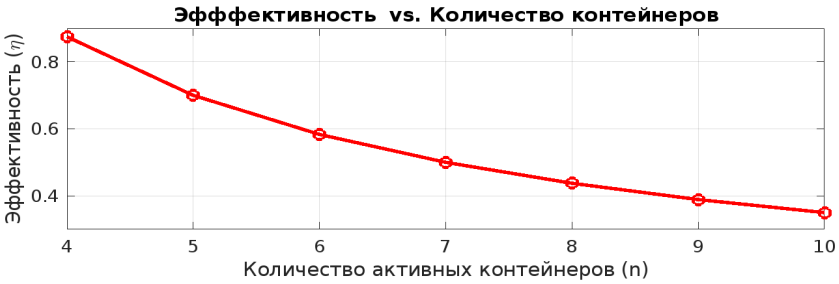

Рис. 5. Эффективность в зависимости от количества контейнеров

Fig. 5. Efficiency depending on the number of containers

Требуется найти оптимальное количество контейнеров, которое сбалансирует высокую эффективность со стабильностью системы. Если η слишком велико (близко к 1), система может испытывать задержки из-за перегрузки. Если η слишком мало, контейнеры тратятся впустую, что приводит к более высоким затратам без реальных преимуществ производительности.

Демонстрация на рис. 6, позволяет утверждать, что с ростом интенсивности трафика (λ) резко увеличивается время отклика ($T_{response}$).

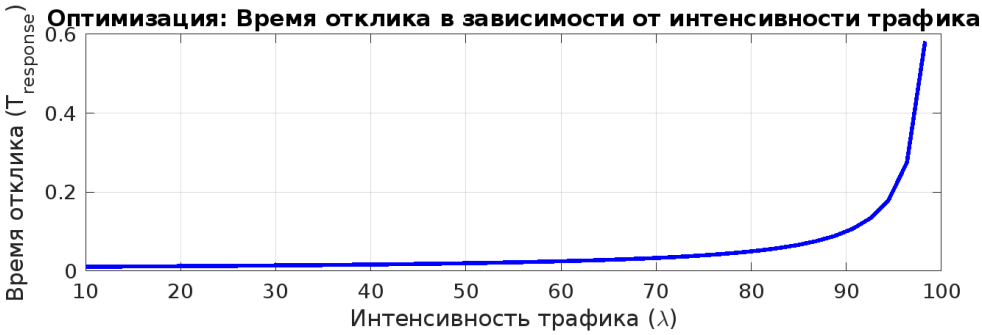

Рис. 6. Оптимизация определением времени отклика в зависимости от интенсивности трафика
Fig. 6. Optimization by determining response time depending on traffic intensity

Когда λ приближается к общей скорости обслуживания (μ), время отклика стремится к бесконечности, что указывает на перегрузку системы. Если интенсивность поступления запросов λ превышает μ, запросы не могут обрабатываться достаточно быстро, что приводит к длинным очередям.

Соответственно, если время отклика слишком велико, не выполняются условия пользовательского опыта и соглашения об уровне обслуживания. Для оптимизации требуется обеспечить $\lambda < \mu$ (т.е. избегать перегрузки системы). Стратегии масштабирования заключаются в динамическом увеличении скорости обслуживания (μ) на основе неопределенности нагрузки.

Функция стоимости (рис. 7) означает наличие точки минимальной стоимости. Если используется слишком мало контейнеров, стоимость высока из-за длительного времени ожидания (плохое обслуживание). Если используется слишком много контейнеров, стоимость увеличивается из-за более высоких эксплуатационных расходов. Стоимость – это комбинированный показатель, который включает в себя прямые затраты на ресурсы (фиксированные и переменные) и неявные затраты (или выгоду) связанные с качеством обслуживания, измеряемым временем ожидания.

Рис. 7. Оптимизация
Fig. 7. Optimization

Формула ожидаемой загрузки (15) $E(\rho)$ описывает среднюю нагрузку на систему в зависимости от распределения вероятностей активных контейнеров. В графике затрат рис. 7 она косвенно используется для расчета потери производительности (через время ожидания), которая, в свою очередь, влияет на общую стоимость системы и оптимальное количество контейнеров. Цель оптимизации состоит в том, чтобы найти такое количество контейнеров (n), при котором эта общая стоимость будет минимальной.

Последствия для оптимизации:

– оптимальное количество контейнеров – это то, где стоимость минимальна;

– недостаточное количество контейнеров приводит к длительным задержкам, что вредит эффективности;

– достаточно много контейнеров тратят ресурсы и неоправданно увеличивают затраты.

Таким образом, разработана новая формула, основанная на теории очередей, которая посредством вычислительной и программной реализации позволяет рассчитывать интенсивность обслуживания за заданный временной интервал и определять оптимальную конфигурацию распределенной контейнерной системы. Результаты демонстрируют способность модели эффективно распределять запросы с учетом ограничений ресурсов и поведения системы с течением времени. Экспериментальное исследование, проведенное в режиме реального времени с использованием *Apache JMeter* для наблюдения за закономерностями трафика в кластере контейнерных систем, подтверждает практическую применимость модели. Балансировщик нагрузки поддерживал стабильное распределение запросов, подтверждая правильность математического подхода к оценке и повышению оптимизации и эффективности распределенных вычислительных систем.

Новизна теоретически обоснованной оптимизации состоит в точном предсказании поведения системы и оптимизации ресурсов, учитывая реалистичную изменчивость нагрузки, а не только ее среднее значение. Математическое моделирование с данными по интенсивности запросов в условиях неопределенности позволяет утверждать, что предложенная модель, интегрирующая вероятностное распределение активных контейнеров ($P(n)$) и стохастические неопределенности нагрузки поступления запросов в единую систему оптимизации, обеспечивает уникальный механизм для адаптивного планирования ресурсов, минимизируя как переизбыток мощностей, так и риски перегрузки, чего не достигают существующие методы, ориентированные на детерминированные или менее динамичные сценарии. Новаторская составляющая состоит в:

– разработке динамической формулы оптимизации в условиях неопределенного трафика;

– интеграции теории массового обслуживания с мониторингом системы в реальном времени для адаптивного управления развертыванием контейнеров;

– применении модели к реальным контейнерным системам и её валидации с использованием многодневных экспериментальных данных.

Заключение

Исследование оптимизации и эффективности распределительной компьютерной системы с контейнерной виртуализацией в условиях неопределенности доказывает возможность создания системы мониторинга и балансировки запросов при внедрении математической модели теории массового обслуживания с адаптивной стратегией, позволяющей балансировать систему распределения и обеспечить оптимальное количество контейнеров, отвечая требованию рентабельности. Выявление взаимосвязанности между метриками интенсивности запросов, времени обслуживания и количества контейнеров позволяет реализовать достижение оптимального количества контейнеров, которое позволяет сбалансировать высокую эффективность и стабильность системы в определенное время суток.

Предложенная математическая модель позволяет поддерживать экономическую эффективность, динамически корректируя количество контейнеров на основе наблюдаемых метрик. Устанавливая взаимосвязь между ключевыми параметрами системы – интенсивностью запросов, временем обслуживания и количеством контейнеров, – модель позволяет разработчикам систем достичь сбалансированного соотношения производительности, стабильности и стоимости, особенно в сценариях с меняющейся рабочей нагрузкой.

Основные результаты работы можно классифицировать как теоретические, так и прикладные:

1) теоретически она способствует использованию моделей очередей в условиях адаптивных ограничений в распределенных системах;

2) на практике она закладывает основу для построения механизмов автоматического масштабирования на основе метрик, получаемых в режиме реального времени;

3) методологически показано, как интегрировать формальные модели с эмпирическими ограничениями для принятия решений в условиях неопределенности.

Перспективой исследования является создание симуляционной экспериментальной модели при различных технических характеристиках для выявления параметра универсализации разработанного математического аппарата к различным распределенным компьютерным системам.

Список источников:

1. Богатырев В.А. Оптимизация резервированного распределения запросов в кластерных системах реального времени // Информационные технологии. – 2015. – Т. 21. – № 7. – С. 495-502.
2. Ефремов А.А., Герасенко К.П., Акулич Р.В., Русина Н.В. Практическое применение распределения Пуассона // ГМинск: 60-я Юбилейная Научная Конференция Аспирантов, Магистрантов и Студентов БГУИР. – 2024. – С. 159-162.
3. Клейнрок Л. Теория массового обслуживания. – М.: Машиностроение, 1979. – С. 112-126.
4. Кузнецов В.В. Системы массового обслуживания. – М.: Юрайт, 2025. – 332 с.
5. Фунг В.К., Богатырев В.А., Кармановский В.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.
6. Плескунов М.А. Теория массового обслуживания: учебное пособие. – Екатеринбург: Издательство Уральского университета, 2022. – 264 с.
7. Apache JMeter. Режим доступа: https://jmeter.apache.org/ (21.11.2024).
8. Crescenzo A., Giorno V., Kumar B.K., Nobile A.G. M/M/1 queue in two alternating environments and its heavy traffic approximation // Journal of Mathematical Analysis and Applications. – 2021 (2018). – pp. 927-1001.
9. Ding S. Analysis and Application of Computer Queueing Theory // Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023). Advances in Computer Science Research. – 2023. – V. 108. – pp. 447-453.
10. Naidu NRV. 06IP/IM74 Operations research. UNIT - 5: Queuing Theory. Bangalore: MSRIT. – 2020. – pp. 45.

References:

1. Bogatyrev V.A. Optimizing Redundant Distribution of the Queries in the Cluster Real-Time Systems. Information Technologies. 2015;21(7): 495-502.
2. Efremov AA, Gerasenko KP, Akulich RV, et al. Practical Application of the Poisson Distribution. In: Proceedings of the 60th Anniversary Scientific Conference of Postgraduates, Master's Students and Students of Belarusian State University of Informatics and Radioelectronics; Minsk; 2024. p. 159-162.
3. Kleinrock L. Queuing Systems. Moscow: Mashinostroenie; 1979. p. 112-126.
4. Kuznetsov V.V. Mass Service Systems. Moscow: Yurayt; 2025.
5. Fung WK, Bogatyrev VA, Karmanskiy VS, et al. Probabilistic and Temporal Characteristics Estimation of a Computer System with Container Virtualization. Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2024;24(2):249-255.
6. Pleskunov M.A. Queuing Theory. Yekaterinburg: Ural University Press; 2022.
7. Apache JMeter [Internet] [cited 2024 Nov 21]. Available from: https://jmeter.apache.org
8. Crescenzo A, Giorno V, Kumar BK, Nobile AG. M/M/1 Queue in Two Alternating Environments and Its Heavy Traffic Approximation. Journal of Mathematical Analysis and Applications. 2018; 465:927-1001.
9. Ding S. Analysis and Application of Computer Queueing Theory. In: Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023). Advances in Computer Science Research. 2023;108:447-453.
10. Naidu NRV. 06IP/IM74 Operations Research. UNIT – 5: Queuing Theory. Bangalore: MSRIT; 2020.

Информация об авторах:

Чудинова Александра Анатольевна
аспирант Университета ИТМО, член Всероссийского сообщества Знание, ORCID 0000-0003-4171-6964, ORJ-2772-2025, SPIN: 7984-9563.

Information about the authors:

Chudinova Alexandra Anatolyevna
Postgraduate Student of ITMO University, member of the All-Russian Knowledge Community, ORCID: 0000-0003-4171-6964, ORJ-2772-2025, SPIN: 7984-9563

Статья поступила в редакцию 26.11.2025; одобрена после рецензирования 14.12.2025; принята к публикации 21.12.2025.

The article was submitted 26.11.2025; approved after reviewing 14.12.2025; accepted for publication 21.12.2025.

Рецензент – Малаханова А.Г., кандидат технических наук, доцент, Брянский государственный технический университет.

Reviewer – Malakhanova A.G., Candidate of Technical Sciences, Associate Professor, Bryansk State Technical University.