УДК 004.75/.052.2

ИССЛЕДОВАНИЕ МОДЕЛЕЙ ОПТИМИЗАЦИИ УПРАВЛЕНИЯ ТРАФИКОМ СЕРВИС-ОРИЕНТИРОВАННЫХ ОБЛАЧНЫХ ПРИЛОЖЕНИЙ В ПРОГРАММНО-УПРАВЛЯЕМОЙ ИНФРАСТРУКТУРЕ ВИРТУАЛЬНОГО ЦЕНТРА ОБРАБОТКИ ДАННЫХ

Болодурина И.П., Парфёнов Д.И.

Оренбургский государственный университет,
Оренбург, e-mail: prmat@mail.osu.ru, fdot_it@mail.osu.ru

В настоящее время доля использования технологии облачных вычислений в современных бизнес-процессах компаний неуклонно растет. Несмотря на то, что это позволяет снижать стоимость владения и эксплуатации ИТ-инфраструктуры, существует ряд проблем связанных с управлением центрами обработки данных. Одной из таких проблем является эффективность использования имеющихся в распоряжении компаний вычислительных и сетевых ресурсов. Одним из направлений оптимизации является процесс управления трафиком сервис-ориентированных облачных приложений в центрах обработки данных (ЦОД). При многозвенной архитектуре современных ЦОД такая задача весьма не тривиальная. Преимуществом современной инфраструктуры виртуализации является возможность использования программно-конфигурируемых сетей и программно-управляемых хранилищ данных. Однако существующие алгоритмические решения при оптимизации не учитывают ряд особенностей формирования трафика в сети с несколькими классами приложений. В рамках проведенного исследования решена задача оптимизации распределения трафика сервис-ориентированных облачных приложений для программно-управляемой инфраструктуры виртуального ЦОД. Предложена имитационная модель, позволяющая описать трафик в программно-конфигурируемых сегментах сети ЦОД, участвующих в обработке запросов пользователей к сервис-ориентированным облачным приложениям, расположенным в сетевой среде, включающей в себя гетерогенную облачную платформу и программно-конфигурируемые хранилища данных. Разработанная модель позволила реализовать алгоритм управления трафиком облачных приложений и оптимизировать доступ к системе хранения, за счет эффективного использования канала для передачи данных. В ходе экспериментальных исследований установлено, что применение разработанного алгоритма позволяет сократить время отклика сервис-ориентированных облачных приложений и, как следствие, повысить производительность обработки запросов пользователей и снизить количество отказов.

Ключевые слова: облачные вычисления, виртуальный центр обработки данных, программно-конфигурируемая инфраструктура, программно-конфигурируемая сеть

RESEARCH MODELS OF TRAFFIC CONTROL TO OPTIMIZE SERVICE-ORIENTED CLOUD APPLICATIONS IN A SOFTWARE-DEFINED INFRASTRUCTURE IN THE VIRTUAL DATA CENTER

Bolodurina I.P., Parfenov D.I.

Orenburg State University, Orenburg, e-mail: prmat@mail.osu.ru, fdot_it@mail.osu.ru

Currently, the proportion of use of cloud computing technology in today's business processes of companies is growing steadily. Despite the fact that it allows you to reduce the cost of ownership and operation of IT infrastructure, there are a number of problems related to the control of data centers. One such problem is the efficiency of the use of available companies compute and network resources. One of the directions of optimization is the process of traffic control of service-oriented cloud applications in data centers. Given the multi-tier architecture of modern data center, this problem does not quite trivial. The advantage of modern virtual infrastructure is the ability to use software-defined networks and software-defined data storages. However, existing solutions with algorithmic optimization does not take into account a number of features forming network traffic with multiple classes of applications. Within the framework of the exploration solved the problem of optimizing the distribution of traffic service-oriented cloud applications for the software-defined virtual data center infrastructure. A simulation model describing the traffic in data center and software-defined network segments involved in the processing of user requests for service-oriented cloud applications located network environment that includes a heterogeneous cloud platform and software-defined data storages. The developed model has allowed to implement cloud applications traffic control algorithm and optimize access to the storage system through the effective use of the channel for data transmission. In experimental studies found that the application of the developed algorithm can reduce the response time of service-oriented cloud applications, and as a result improve the performance of processing user requests and to reduce the number of failures.

Keywords: cloud computing, virtual data center, software-defined infrastructure, software-defined network

В настоящее время доля использования облачных сервис-ориентированных приложений для автоматизации современных бизнес процессов в средних и крупных компаниях неуклонно растет. Это позволяет снижать стоимость владения и эксплуатации ИТ-инфраструктуры за счет более эффективного использования имеющихся в распоряжении компании вычислительных и сетевых ресурсов. На сегодняшний день решения, применяемые при построении виртуальной инфраструктуры, динамично

развиваются. Так в последнее время для размещения облачных сервис-ориентированных приложений в инфраструктуре виртуального центра обработки данных (ЦОД) активно применяется технология контейнеров. Наиболее востребованным подходом является использование контейнеризации приложений и сервисов на базе Docker. Кроме того, современные ЦОД переходят от физической инфраструктуры к виртуальной с применением программно-конфигурируемых компонентов (сетей, хранилищ данных и др.) [2, 3]. Это вносит свои коррективы в механизмы управления запуском и размещением сервис-ориентированных облачных приложений.

Наиболее существенные изменения происходят в задачах планирования SaaS- и PaaS-сервисов [4]. При этом традиционные способы решения задач планирования и управления распределением ресурсов, количественно максимизирующие выполнение запросов пользователей, применяемые в высокопроизводительных (High Performance Computing, HPC) системах, оказываются неэффективны [1, 5]. В основном это связано с тем, что для HPC время отклика на выполнение запроса пользователя при решении задач из пакета запросов играет значительно меньшую роль, нежели сам факт выполнения запроса. При этом в HPC задачах чаще всего делается акцент на скорость выполнения на различных конфигурациях вычислительных систем. Для сервис-ориентированных облачных приложений, наоборот, наиболее критичным параметром является время отклика на запрос пользователя [6–7].

В рамках настоящего исследования построена имитационная модель программно-управляемой инфраструктуры виртуального ЦОД, позволяющая описать трафик запросов пользователей к сервис-ориентированным облачным приложениям в гибко реконфигурируемой виртуальной сетевой среде, включающей в себя гетерогенную облачную платформу и программно-конфигурируемые хранилища данных.

Как правило, в программно-управляемой инфраструктуре виртуального ЦОД размещаются несколько неоднородных сервис-ориентированных облачных приложений. На основании этого предположим, что в сети виртуального ЦОД присутствуют как минимум три класса трафика приложений, таких как веб-приложения; case-приложения (прикладное ПО, доступное по DaaS или SaaS модели); видеосервисы. При этом в качестве трафика приложений будем рассматривать запросы пользователей к каждому классу приложений. Для генерации запросов пользователей в имитационной модели применим к каждому классу трафика весовые коэффициенты k_1, k_2, k_3. Каждый из перечисленных коэффициентов позволяет разделить заявки на классы и оказывает влияние на следующий набор параметров: время выполнения, маршрут в имитационной модели, приоритет в очереди на обработку, интенсивность поступления, а также закон распределения, согласно которому осуществляется генерация трафика определенного класса.

Представим имитационную модель программно-управляемой инфраструктуры виртуального ЦОД в виде многоканальной системы массового обслуживания. В ее состав входит источник заявок пользователей (I), очередь (Q_s) и планировщик (S), управляющий процессами размещения и запуска приложений и сервисов (App), а также пул вычислительных узлов (Srv) и систем хранения (Stg) ЦОД, содержащие как сами приложения, так и требуемые имм данные. Схема СМО программно-управляемой инфраструктуры виртуального ЦОД представлена на рис. 1.

Модель СМО носит стохастический характер. Для ее работы необходимо создать поток запросов пользователей к облачным приложениям и сервисам, учитывающий законы распределения и интенсивность поступления заявок для каждого класса облачного приложения и сервиса.

Для решения задачи по оптимизации управления размещением сервис-ориентированных приложений в облачной среде виртуального ЦОД необходимо определить законы распределения трафика для каждого класса приложений, а также распределить сам трафик по объектам доступа (виртуальным серверам, контейнерам и системам хранения). Для этого необходимо установить определенный маршрут и построить для него закон управления на временном интервале $T = [t_1, t_2]$.

В динамике трафик облачных приложений и сервисов в программно-управляемой инфраструктуре виртуального ЦОД можно описать следующей дискретной системой:

$$x_{i,j}(t+\Delta t) = x_{i,j}(t) - \sum_{k=1}^{K}\sum_{l=1}^{N} s_{i,j}(t)u_{i,l}^{j,k}(t) + \sum_{m=1}^{N} s_{m,i}(t)u_{m,l}^{j}(t) + y_{i,j}(t), \quad (1)$$

где N – количество виртуальных узлов в сети; K – количество классов приложений и сервисов в сети; $s_{i,j}(t)$ – пропускная способность каналов между i-м вычислительным узлом и j-й системой хранения данных ($i \neq j$); $y_{i,j}(t) = \lambda_{i,j}(t)\Delta t$ – объем трафика (количество запросов пользователей), поступающий в момент времени t на виртуальный узел i и предназначенной для передачи системе хранения j; $\lambda_{i,j}(t)$ – интенсивность поступающей нагрузки, которая определяется как суммарная интенсивность потока запросов пользователей, подключаемых к виртуальному узлу i и ведущих обмен с j-м узлом системы хранения данных; $u_{i,l}^{j,k}(t)$ – доля пропускной способности канала, выделенного в сегменте программно-конфигурируемой сети (i, l), в момент времени t потоку запросов пользователей к приложению k-го класса, осуществляющего работу с данными в системе хранения j.

Чтобы исключить возможность перегрузки объектов виртуального ЦОД, ввиду ограниченности буферов очередей на вычислительных узлах, а также эффективного использования пропускных способностей каналов передачи данных на переменные, отвечающие за управление и формирование канала для облачных приложений в программно-конфигурируемой сети (ПКС) накладывается ряд ограничений.

Ограничения на переменные динамического управления сетевыми ресурсами связаны с ограниченностью физической пропускной способности каналов между сетевыми устройствами и могут быть записаны в следующем виде:

$$0 \le u_{i,l}^{j}(t) \le u_{i,l}^{j(\max)} \le 1; \quad \sum_{l=1}^{N} u_{i,l}^{j}(t) \le \varepsilon_{i,l}^{j,k} \le 1, \quad (2)$$

где $u_{i,l}^{j(\max)}$ – максимальный предел выделяемой доли пропускной способности доступный вычислительному узлу i в сегменте ПКС l для передачи трафика к системе хранения j; $\varepsilon_{i,l}^{j,k}$ доля пропускной способности канала для вычислительного узла i в сегменте ПКС l, выделенная для передачи пакетов запросов пользователей к приложению k-го класса для реализации динамической стратегии управления виртуальными ресурсами при доступе к системе хранения j.

Организация программно-управляемой инфраструктуры виртуального ЦОД с применением программно-конфигурируемой сети позволяет управлять механизмами формирования очередей, что вводит дополнительные ограничения на переменные состояния

$$0 \le x_{i,j}(t) \le x_{i,j}^{(\max)}; \quad \sum_{l=1}^{N} x_{i,j}(t) \le x_{i}^{(\max)}, \quad (3)$$

где $x_{i,j}^{(\max)}$ – максимально допустимая длина очереди на i-м вычислительном узле x, выделенная для обработки поступающего трафика к системе хранения j; $x_{i}^{(\max)}$ максимально допустимый объем буфера на узле вычислительном узле i.

В качестве критерия оптимальности рассмотрим максимизацию производительности системы, достигаемой за фиксированный период $T = [t_1, t_2]$, которая в рамках модели формализуется в виде целевой функции вида

$$\sum_{t=0}^{t-1}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N} s_{i,j}(t)u_{i,l}^{j,k}(t) \to \max. \quad (4)$$

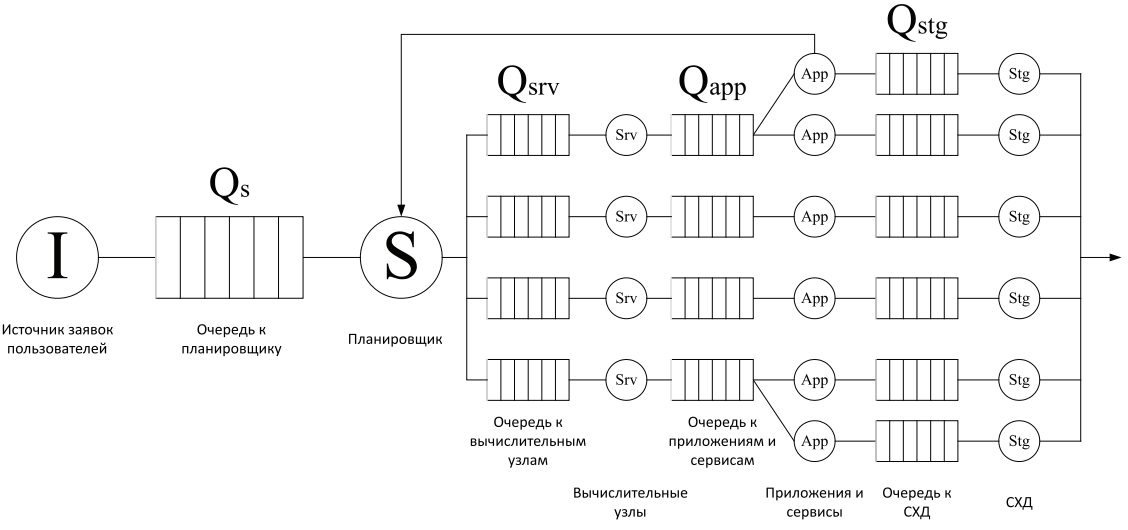

Рис. 1. Схема СМО программно-управляемой инфраструктуры виртуального ЦОД

Для решения задачи оптимизации применим итерационный метод, который позволяет исследовать динамику работы системы на интервале $T = [t_1, t_2]$ и осуществлять управление пропускной способностью канала для выделенного класса приложений в программно-конфигурируемой сети.

Алгоритм управления трафиком облачных сервис-ориентированных приложений в программно-управляемой инфраструктуре виртуального центра обработки данных

Для решения оптимизационной задачи нами разработан алгоритм управления трафиком облачных сервис-ориентированных приложений в программно-управляемой инфраструктуре виртуального центра обработки данных. По сравнению с имеющимися аналогами алгоритм использует эвристический анализ потоков запросов, а также классификацию трафика по типу передаваемых данных в процессе работы приложения, размещенного в облачной платформе. Гибкость предлагаемого решения обусловлена виртуализацией хранилища данных. Это позволяет динамически изменять расположение приложений в облачной системе относительно физических устройств, что дает возможность предоставлять непрерывный доступ к услугам и сервисам. Предлагаемое решение прозрачно для клиента и масштабирует облачные приложения на несколько виртуальных устройств хранения. Это обеспечивает сокращение времени отклика приложения, а также повышает отказоустойчивость всей системы в целом. Формирование программно-управляемых самоорганизующихся хранилищ данных на базе виртуальных машин и контейнеров позволяет не только снизить риски, связанные с потерей или недоступностью данных, но и обеспечивает интеллектуальный анализ востребованности облачных приложений. На базе полученных данных формируются карты размещения виртуальных машин и контейнеров, а также правила для формирования потоков в программно-конфигурируемой сети виртуального ЦОД. В основу алгоритма управления трафиком облачных приложений и сервисов в программно-управляемой инфраструктуре положена разработанная имитационная модель. Полученная информация о характере распределения и интенсивности поступления запросов подвергается анализу с применением алгоритмов машинного обучения (Data Mining), основанного на нейросетевом подходе. В результате формируется карта оптимального расположения приложений и сервисов внутри самой облачной платформы с привязкой к физическим устройствам, а также создается карта маршрутов для формирования потоков трафика с учетом востребованности данных в системе хранения. Путем анализа двух карт и эвристического алгоритма прогнозирования системы управления виртуальным ЦОД принимается решение о реконфигурации структуры облачной платформы и перестройке маршрутов для выделенных классов трафика. При этом обе карты являются динамическими объектами, формируемыми не только по мере возникновения событий в программно-управляемой инфраструктуре, но и с заданным интервалом времени Δt, подбираемым индивидуально для каждой облачной платформы. В рамках исследования определен наиболее оптимальный интервал времени для анализа и перестроения карт, при котором работа системы будет наиболее эффективной.

При работе с облачными приложениями и сервисами не исключена ситуация, при которой для обслуживания запроса пользователя могут быть задействованы сразу несколько типов ресурсов инфраструктуры виртуального ЦОД с различными характеристиками доступа. При работе с такими данными облачной системе необходимо осуществлять подготовку инфраструктуры доступа для оптимизации времени отклика на запрос. Для этого разработанный алгоритм управления трафиком облачных приложений и сервисов в программно-управляемой инфраструктуре виртуального ЦОД в ходе работы строит план выполнения запросов, тем самым подстраивая каждый задействованный объект инфраструктуры под поток запросов пользователей. В результате план выполнения потока запросов пользователей с одинаковой интенсивностью в разные моменты времени может быть составлен поразному. Перестроение плана происходит в соответствии с востребованностью ресурсов, что позволяет эффективно управлять распределением и динамической балансировкой нагрузки в программно-управляемой инфраструктуре виртуального ЦОД. Обобщенная блок-схема работы алгоритма управления трафиком облачных приложений и сервисов в программно-управляемой инфраструктуре виртуального центра обработки данных представлена на рис. 2.

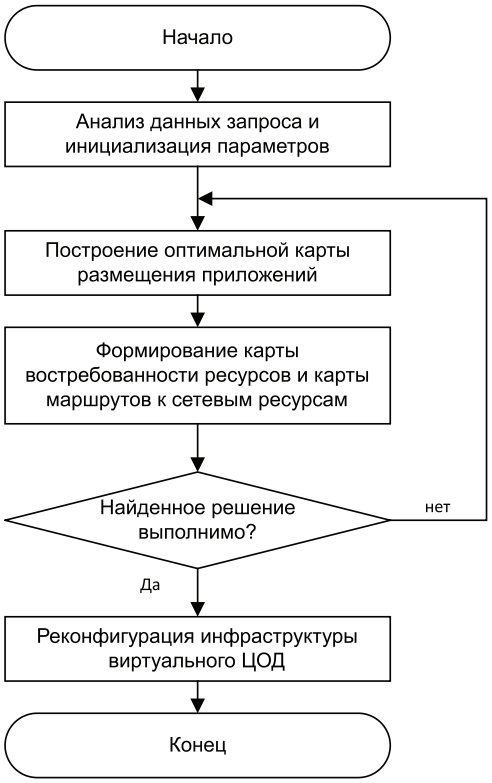

Рис. 2. Обобщенная блок-схема работы алгоритма управления трафиком облачных сервис-ориентированных приложений в программно-управляемой инфраструктуре виртуального центра обработки данных

Всего в работе алгоритма управления трафиком облачных сервис-ориентированных приложений в программно-управляемой инфраструктуре виртуального центра обработки данных можно выделить три этапа планирования и оптимизации, выполняемых для обслуживания запросов пользователей.

На первом этапе в плане выполнения анализируются характеристики запроса, поступающего от пользователя к приложению, а именно определяется тип приложения и вид передаваемых данных. Запрос данных является многокомпонентным, т.е. для организации канала задействуется сразу несколько элементов системы. Поэтому на втором этапе система управления определяет наиболее подходящие ресурсы, способные гарантировать выполнение запроса. При этом если среди ресурсов существует вариативность, то на основе информации о прогнозе востребованности каждого из доступных вариантов ресурса строится ранжированный список с привязкой к виртуальному хранилищу данных. При этом при построении маршрута трафика учитываются только наименее загруженные хранилища данных и каналы связи. На третьем этапе работы алгоритма анализируются текущие параметры ресурсов и прогнозируется время, необходимое на выполнение текущего запроса. Полученные результаты сохраняются для дальнейшего использования при самообучении алгоритма. В случаях высокой ресурсоемкости запроса для предварительной оценки в алгоритме применяется поэлементный анализ объектов доступа, входящих в запрос.

Экспериментальные исследования

Для оценки эффективности разработанного алгоритма управления трафиком облачных сервис-ориентированных приложений в программно-конфигурируемых хранилищах данных виртуального ЦОД, нами проведено исследование работы облачной системы, построенной на базе Openstack с различными параметрами. Для сравнения в эксперименте использовались алгоритмы, применяемые в облачных системах для управления запуском приложений и размещением данных. Для анализа эффективности и производительности работы алгоритмов на различных системах хранения, нами определены типовые условия эксперимента, включающие традиционные устройства на магнитных дисках (HDD), твердотельные накопители (SSD) и виртуальные хранилища (SDS). Для экспериментального исследования создан прототип облачной среды, включающий в себя основные узлы, а также программные модули для разработанных алгоритмов, выполняющие перераспределение запросов пользователей к данным в программно-конфигурируемом хранилище.

В облачной системе OpenStack реализован модуль, применяющий разработанный алгоритм управления трафиком облачных приложений и сервисов в программно-конфигурируемых хранилищах данных виртуального ЦОД для рационального использования вычислительных ресурсов облачной системы и эффективного размещения виртуальных машин по физическим узлам, а также связанных с ними данных. В ходе эксперимента для анализа данных создан поток запросов, аналогичный реальным запросам пользователей к облачным приложениям, основанный на данных лог-записей доступа к определенным видам ресурсов с классификацией по типам приложений и структуре запроса. Ретроспектива воспроизводимых запросов составила 3 года, при этом для нагрузочного эксперимента

применялись усредненные данные. Полученные данные распределены на пул виртуальных машин по следующим критериям: тип клиента, осуществившего обращение к данным, тип сервиса, востребованного при подключении. При этом количество одновременных запросов, поступивших в систему, составило 100000, что соответствует максимальному числу потенциальных пользователей системы.

Все сформированные запросы воспроизводились последовательно на трех экспериментальных площадках. Данное ограничение введено в связи с необходимостью сопоставления результатов с физическими системами хранения данных, не способных к реконфигурации. Основным отличием экспериментальных площадок является использование твердотельных накопителей.

Помимо площадок для анализа эффективности сформировано 3 группы экспериментов, направленных на интенсивное выполнение запросов по чтению (эксперимент 1), записи (эксперимент 2) и одновременных операциях чтения и записи данных (эксперимент 3) для каждого класса приложения.

Время эксперимента составило один час, что соответствует наиболее длительному периоду времени пиковой нагрузки системы, зафиксированному в реальном трафике. Анализ данных экспериментальных исследований доказал, что управление трафиком облачных приложений и сервисов в программно-конфигурируемых хранилищах данных виртуального ЦОД с использованием предложенного алгоритма более эффективно независимо от типа физических устройств, используемых в хранилище данных.

Полученные данные подтверждают целесообразность применения разработанного алгоритма для предоставления эффективного доступа к сервис-ориентированным приложениям облачных систем. При этом по результатам проведенных экспериментов также можно сделать вывод о снижении на 20–25 % количества отказов при обслуживании запросов приложениями и сервисами, размещенными в программно-управляемых хранилищах данных, позволяющих гибко адаптировать инфраструктуру в зависимости от нагрузки.

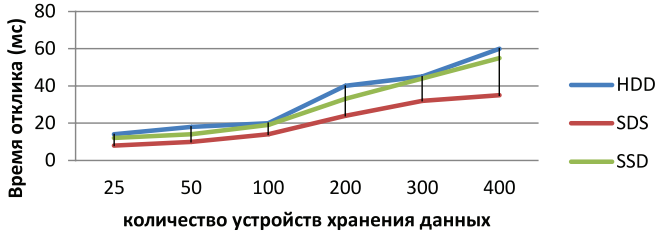

Рис. 3. Анализ времени отклика приложений при выполнении запросов на чтение данных (эксперимент 1)

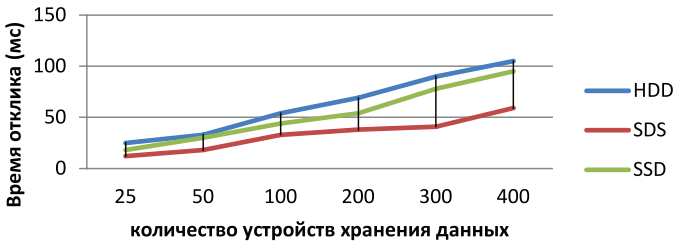

Рис. 4. Анализ времени отклика приложений при выполнении запросов на запись данных (эксперимент 2)

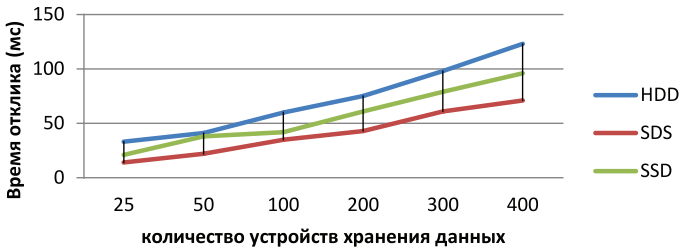

Рис. 5. Анализ времени отклика приложений при одновременном выполнении запросов на чтение и записи данных (эксперимент 3)

Выводы

В результате проведенного исследования решена задача оптимизации распределения трафика облачных сервис-ориентированных приложений для программно-управляемой инфраструктуры виртуального ЦОД. Предложена имитационная модель, позволяющая описать трафик в программно-конфигурируемых сегментах сети ЦОД, участвующих в обработке запросов пользователей к приложениям и сервисам расположенных в сетевой среде, включающей в себя гетерогенную облачную платформу и программно-конфигурируемые хранилища данных. Разработанная модель позволила реализовать алгоритм управления трафиком облачных приложений и оптимизировать доступ к системе хранения, за счет эффективного использования канала для передачи данных.

В ходе экспериментальных исследований установлено, что применение разработанного алгоритма позволяет сократить время отклика облачных приложений и сервисов и, как следствие, повысить производительность обработки запросов пользователей и снизить количество отказов.

Работа выполнена при поддержке РФФИ (научные проекты 16-37-60086 мол_а_дк и 16-07-01004), Президента Российской Федерации, грант для государственной поддержки молодых российских ученых – кандидатов наук (МК-1624.2017.9).

Список литературы

1. Болодурина И.П., Парфёнов Д.И. Алгоритмы комплексной оптимизации потребления вычислительных ресурсов в облачной системе дистанционного обучения / И.П. Болодурина, Д.И. Парфёнов // Вестник Оренбургского государственного университета. – 2013. – № 9. – С. 177–184.

2. Болодурина И.П., Парфёнов Д.И. Управление потоками данных в высоконагруженных информационных системах, построенных на базе облачных вычислений // Системы управления и информационные технологии. – 2015. – № 1.1. – С. 111–118.

3. Bocchi E., Drago I., Mellia M. Personal Cloud Storage Benchmarks and Comparison // IEEE Transactions on Cloud Computing. – 2015. – Vol. 99. – IEEE, 2015. – P. 1–14.

4. Bolodurina I., Parfenov D., Shukhman A. Approach to the effective controlling cloud computing resources in data centers for providing multimedia services // Control and Communications (SIBCON), 2015 International Siberian Conference on. – IEEE, 2015. – P. 1–6.

5. Charuenporn P., Intakosum S. Qos-Security Metrics Based on ITIL and COBIT Standard for Measurement Web Services // J. UCS. – 2012. – Vol. 18. – № 6. – P. 775–797; available at: www. http://jucs.org/jucs_18_6.

6. Rajiv R., Benatallah B., Schahram D., Michael P. Cloud Resource Orchestration Programming: Overview, Issues, and Directions // IEEE Internet Computing. – 2015. – Vol. 19, Issue: 5. – P. 46–56.

7. Thiago A.L., Genez L.F., Bittencourt E., Madeira R.M. Workflow scheduling for SaaS / PaaS cloud providers considering two SLA levels // Network Operations and Management Symposium (NOMS). – IEEE, 2012. – P. 906–912.