УДК 519.876.5:004.946

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ АНАЛИТИЧЕСКОГО МОДЕЛИРОВАНИЯ ДЛЯ ОЦЕНКИ ПРОИЗВОДИТЕЛЬНОСТИ ОБЛАЧНЫХ ПРИЛОЖЕНИЙ

В.В. Гайдамако

Оценка производительности облачных приложений и сервисов важна как для поставщика облака, так и для клиента, она производится как на этапе проектирования, так и во время эксплуатации системы для мониторинга и обеспечения гарантированного качества обслуживания. Облачные приложения разворачиваются в динамической среде, экземпляры сервиса могут добавляться, удаляться, мигрировать, что усложняет оценку производительности. Рассмотрены методы аналитического моделирования, основанные на теории массового обслуживания, теории управления, сетевом исчислении, исчислении реального времени. Все эти методы могут применяться для создания моделей приложений и облачной инфраструктуры, но следует отметить, что сетевое исчисление и исчисление реального времени представляются наиболее перспективными для оценки производительности облачных сервисов, так как они дают гарантии жесткого реального времени.

Ключевые слова: оценка производительности; аналитическое моделирование; сетевое исчисление; исчисление реального времени; теория управления; теория массового обслуживания; стохастические сети с вознаграждением.

БУЛУТ ТИРКЕМЕЛЕРИНИН ӨНДҮРҮМДҮҮЛҮГҮН БААЛОО ҮЧҮН АНАЛИТИКАЛЫК МОДЕЛДӨӨ ЫКМАЛАРЫНА САЛЫШТЫРМА ТАЛДОО ЖҮРГҮЗҮҮ

В.В. Гайдамако

Булут тиркемелеринин жана сервистеринин өндүрүмдүүлүгүн баалоо булут менен камсыз кылуучу үчүн да, кардар үчүн да маанилүү, талдоо жүргүзүү долбоорлоо этабында да, системаны эксплуатациялоо учурунда да тейлөөнүн кепилденген сапатын камсыз кылуу жана мониторинг жасоо үчүн жүргүзүлөт. Булут тиркемелери динамикалык чөйрөдө ишке ашырылат, сервистин нускамалары кошулуп, өчүрүлүп, миграцияланууу мүмкүндүгүнө ээ, мунун өзү өндүрүмдүүлүктү баалоону татаалдаштырат. Макалада массалык тейлөөгө, башкаруу теориясына, тармактык эсептөөгө, реалдуу убакытты эсептөөгө негизделген аналитикалык моделдөө ыкмалары каралды. Ушул ыкмалардын бардыгы тиркемелердин моделдерин жана булут инфратүзүмүн түзүүдө колдонулушу мүмкүн, бирок булут кызматтарынын өндүрүмдүүлүгүн баалоо үчүн тармактык эсептөө жана реалдуу убакытты эсептөө бир кыйла келечектүү болуп эсептеле тургандыгын белгилей кетүү керек, анткени алар реалдуу убакытка кепилдик беришет.

Түйүндүү сөздөр: өндүрүмдүүлүктү баалоо; аналитикалык моделдөө; тармактык эсептөө; реалдуу убакытты эсептөө; башкаруу теориясы; массалык тейлөө теориясы; сыйлык менен стохастическалык тармактар.

COMPARATIVE ANALYSIS OF ANALYTICAL MODELING METHODS FOR EVALUATING THE PERFORMANCE OF CLOUD APPLICATIONS

V.V. Gaidamako

Performance evaluation the of cloud applications and services is important for both the cloud provider and the client, both at the design stage and during the operation of the system to monitor and ensure quality of service. Cloud applications are deployed in a dynamic environment, service instances can be added, removed, migrated, which complicates performance assessment. Methods of analytical modeling based on queuing theory, control theory, network calculus, real-time calculus are considered. All of these methods can be used to create models of applications and cloud infrastructure, but it should be noted that network and real-time calculus seem to be the most promising for assessing the performance of cloud services, since they provide hard real time guarantees.

Keywords: performance evaluation; analytical modeling; network calculus; real-time calculus; control theory; queuing theory; reward stochastic networks.

Введение. Оценка производительности облачных приложений крайне важна как при проектировании отдельных приложений и облачной инфраструктуры в целом, так и в процессе эксплуатации – для обеспечения гарантированного качества предоставляемых услуг. Облачные приложения разворачиваются в динамической среде – виртуальных машинах (ВМ), контейнерах, микросервисах, являются многоуровневыми, многоконтейнерными, распределенными по множеству физических серверов, экземпляры сервиса могут добавляться, удаляться, мигрировать, что делает оценку производительности достаточно непростой задачей. Подходы к оценке производительности облачных услуг можно разделить на две или даже три большие группы – оценка, основанная на измерениях, оценка, основанная на аналитическом моделировании и оценка, основанная на имитационном моделировании [1, 2]. На этапе проектирования предпочтительным подходом является аналитическое моделирование как наименее затратное. Методы аналитического моделирования позволяют оценить влияние большого количества параметров на производительность системы еще на стадии планирования и разработки системы и сервисов. В любом случае аналитическое моделирование должно обеспечить адекватность модели моделируемой системе и короткое время анализа, так, чтобы этап моделирования стал частью процесса разработки системы.

Метрики (показатели) оценки производительности. На производительность облачных приложений влияет множество факторов: вычислительная мощность, коммуникации, и доступ к системам хранения. В таблице 1 приводятся основные метрики производительности облачных приложений.

Более подробно метрики производительности описаны в [3]:

Аналитические методы оценки производительности

Теория Массового Обслуживания (ТМО) (Queueing Theory) является классическим подходом к моделированию и анализу компьютерных систем, разработано множество моделей сети массового обслуживания для облачных приложений и центров обработки данных [1, 2, 4]. С помощью ТМО можно оценить среднее время отклика системы (время ожидания + время обслуживания), распределение количества запросов в очереди и в системе, среднюю длину очереди. Для создания адекватной модели важно описать рабочую нагрузку, дисциплину обслуживания, характеристики очереди. В большинстве исследований предполагалось, что процесс поступления запроса на обслуживание является пуассоновским процессом с экспоненциально распределенным временем между поступлениями заявок (inter-arrival time), распределение времени обслуживания принималось экспоненциальным или произвольным. Однако большое разнообразие приложений, использующих разные облачные сервисы, может генерировать рабочие нагрузки с разными шаблонами. Поэтому разработка адекватных моделей для рабочих нагрузок облачных сервисов, точно представляющих трафик, генерируемый широким спектром приложений, является открытой проблемой. Например, в облачных информационно-измерительных системах (ИИС) [5] рабочая нагрузка генерируется физическими датчиками, с одной стороны, и пользователями – с другой, для их описания должны использоваться разные шаблоны. Разработка моделей для рабочих нагрузок от датчиков и пользователей является одной из задач моделирования облачных ИИС.

Методы ТМО обычно применяются для анализа конкретной облачной среды с конкретными допущениями по реализации сервисов. Например, для облачной системы с одним центром обработки данных (ЦОД) могут использоваться различные модели очередей. Очередь M/M/1 [6] – базовая модель ТМО, система с одним сервером, поступление заявок на обслуживание описывается процессом Пуассона, а время обслуживания имеет экспоненциальное распределение. Модель с учетом виртуализации рассматривает приложения как очереди, а виртуальные машины – как серверы, время между поступлением заявок и время обслуживания имеет экспоненциальное распределение, это очередь M/M/m/N – система с m серверами и конечным буфером длиной N [7]. В других работах предложены модели с учетом эластичности – изменения количества серверов в зависимости от длины очереди, совместного использования ресурсов виртуальными машинами, для приложений с приоритетами – модель очереди M/M/m/m [8] с m серверами без буфера, время прибытия – пуассоновский процесс,

Таблица 1 –Типичные метрики, используемые при оценке производительности облачных сервисов

Метрика	Описание
Время отклика сервиса (задержка)	Задержка (время) между запросом на сервис и завершением сервиса
Пропускная способность сервиса	Количество заданий в единицу времени, выполняемое провайдером сервиса
Доступность сервиса	Вероятность принятия запроса провайдером сервиса
Использование системы (коэффициент загрузки системы)	Процент системных ресурсов, требуемых для предоставления сервиса
Устойчивость системы	Стабильность производительности системы во времени, особенно при импульсных нагрузках
Масштабируемость системы	Способность системы сохранять производительность при росте нагрузки из-за увеличения размера или объема системы
Эластичность системы	Способность системы адаптироваться к изменениям нагрузки

время обслуживания – экспоненциальное распределение. Предположение об экспоненциальном времени обслуживания упрощает моделирование, но слишком далеко от реальной картины, поэтому рассматриваются очереди с произвольным распределением времени обслуживания M/G/m. В работе [9] рассматривается модель облачного Центра Обработки Данных (ЦОД) M/G/m/m + r, процесс поступления заявок пуассоновский, заявки обслуживаются в порядке поступления.

Одной из особенностей облачных систем является возможность различий в предоставлении одного и того же сервиса одного провайдера различным пользователям. Сервис может быть в разных ЦОД, на серверах разных типов, он может мигрировать со временем на другой сервер и даже в другой ЦОД. Принцип SOA (Service Oriented Architecture, сервис-ориентированная архитектура) в предоставлении облачных услуг делает реализацию сервиса прозрачной для конечного пользователя, ему не известны детали реализации. Все это делает недопустимыми и неверными любые предположения об определенной архитектуре и технологии реализации облачной системы и не допускает оценку производительности с точки зрения пользователя [1].

Сетевое исчисление (СИ) (Network calculus) [10, 11] использует математический аппарат мини-плюс алгебры, оно сделало возможным применение подхода, основанного на профилировании. Основная идея этого подхода заключается в том, чтобы моделировать и анализировать информацию, характеризующую качество обслуживания (Quality of Service – QoS), а не какие-либо конкретные конфигурации сервисов и рабочей нагрузки. Информация о качестве обслуживания – значения выбранных показателей (метрик) производительности, которые обычно описываются в Соглашении об уровне обслуживания (Service License Agreement – SLA). Обычно в этом соглашении оговаривается минимальная гарантируемая поставщиком облака пропускная способность сервисов и максимальная рабочая нагрузка, которую может предоставить пользователь. Если разработать профили для минимальной пропускной способности, гарантируемой поставщиком услуг, и максимальной рабочей нагрузки, генерируемой пользователем, то на их основе можно будет получить некоторые границы для показателей производительности, например, наихудшую задержку обслуживания и максимальную длину очереди запросов на обслуживание [10]. Два ключевых понятия в сетевом исчислении – это кривая поступления запросов и кривая обслуживания. Кривая обслуживания – это функция времени, которая дает нижнюю границу пропускной способности, которую сервер предоставляет клиенту. Точно так же кривая поступления запросов в сетевом исчислении является функцией времени, которая определяет максимальный объем рабочей нагрузки, которую пользователь может сгенерировать в течение произвольного интервала времени. Используя кривые поступления и обслуживания, сетевое исчисление позволяет определить верхнюю границу задержки любого запроса на обслуживание и максимальной длины очереди запросов на сервере [10, 11].

Исчисление реального времени (ИРВ) (Real-Time Calculus) изначально использовалось для анализа и оптимизации параметров архитектуры систем обработки потоков пакетов [12, 13], как и сетевое исчисление, использует мини-плюс алгебру. В работе [4] обосновано применение методов исчисления реального времени к оценке времени отклика приложений для облачных систем, а также возможности моделирования облачной среды и ее компонент – рабочей нагрузки, обработки задач, выделения ресурсов для виртуальных машин (ВМ), взаимного влияния (интерференции) ВМ на производительность, управления автономными ресурсами, консолидации серверов, а также стратегии масштабирования облачных вычислений (горизонтального и/или вертикального). ИРВ используется для определения и предсказания соблюдения жестких ограничений времени исполнения в системах реального времени. Для проведения модульного анализа производительности с ИРВ (МА-ИРВ) необходимо построить абстрактную модель производительности, включающую модель нагрузки, модель обслуживания и модель обработки (processing). Модель должна включать сведения о доступных вычислительных и коммуникационных ресурсах, о приложениях (или выделенных программных/аппаратных компонентах), а также об архитектуре самой системы. Фреймворк ИРВ состоит из модели задачи, модели ресурсов и исчисления (ИРВ), что позволяет рассматривать потоки событий и их обработку. Реальная система представляется в виде абстрактных аналитических компонент (ИРВ-компоненты), поведение которых может быть детерминированным или недетерминированным. Цель модульного моделирования ИРВ – предсказать время отклика приложения и определить, будет ли оно меньше дедлайна при известных характеристиках работы изолированного приложения на каждом уровне.

В ИРВ ресурсная модель собирает информацию о доступных ресурсах обработки различной аппаратуры, включенной в обработку запросов и возможного картирования функций обработки на этих ресурсах (распределения функций обработки по этим ресурсам). Аналитический фреймворк также рассматривает характеристики потока событий, входящих в систему (то есть запросов клиентов), которые описываются кривыми их поступления. Таким образом, для данной инфраструктуры центра обработки данных, ИРВ исчисление может быть применено для аналитического определения таких свойств, как максимальная задержка входного потока, принимая во внимание дисциплины обслуживания на различных ресурсах обработки.

Значение входных параметров аналитической модели, необходимых для построения отрезков линий для кривых поступления и обслуживания (математических функций), может быть получено прямым измерением на реальных системах, трассировкой, в результате имитационного моделирования.

ИРВ принадлежит к классу так называемых детерминированных теорий массового обслуживания. Он детерминирован в том смысле, что всегда можно найти жесткие верхние и нижние границы показателей производительности (время отклика).

Стохастические сети с вознаграждениями (ССВ) (Stochastic Reward Net – SRN) часто применяются совместно с методами ТМО для оценки производительности облачных сервисов [1]. ССВ по существу являются дополненными стохастическими сетями Петри (Stochastic Petri Nets – SPN) с возможностью определения выходных показателей в качестве наградных функций для оценки производительности сложных систем. В [14] авторы применили стратегию подмодели для упрощения моделирования и анализа крупномасштабных систем облачных вычислений IaaS (Инфраструктура-как-сервис). Для трех основных этапов предоставления услуг – предоставление ресурсов, подготовка ВМ и выполнение – разрабатываются подмодели на основе марковских цепей с непрерывным временем (Continuous-Time Markov Chain CMTC). Затем разрабатывается монолитная модель для представления взаимодействия между тремя этапами с использованием ССВ (SRN), все подмодели объединяются для получения общих результатов производительности облачной службы.

Теория управления (ТУ) – проектируемая система рассматривается как система с обратной связью, один из основных типов систем управления (рисунок 1). Эти системы предназначены для автоматического достижения и поддержания желаемого выходного состояния путем сравнения его с фактическим состоянием.

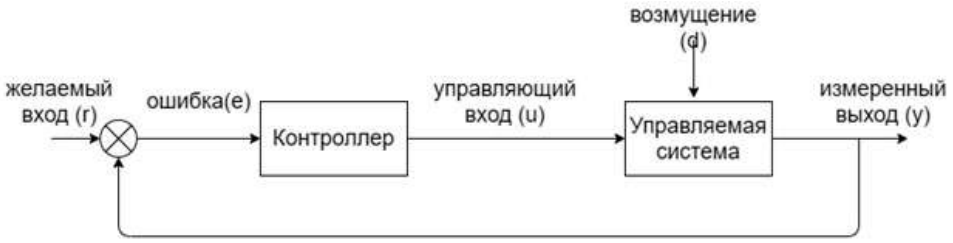

Рисунок 1 – Система с обратной связью [4]

Целевая система предоставляет набор переменных производительности, называемых измеренными выходами или просто выходами. Датчики контролируют выходы целевой системы, а исполнительные механизмы могут настраивать управляющие входы или просто входы для изменения поведения системы. Такие системы обеспечивают устойчивость и быстрое достижение желаемых значений управляемой переменной (например, времени отклика), поэтому используются также для автоматизации масштабирования инфраструктуры облачного приложения в соответствии с динамикой рабочей нагрузки [1, 15].

Блоком принятия решений в системе управления с обратной связью является контроллер обратной связи. Основная задача контроллера – поддерживать выходные параметры системы близко к желаемым значениям, регулируя входы при возмущениях. Это желаемое значение преобразуется системой управления в сигналы уставки, что дает возможность разработчику системы управления определять цели или значения выходов, которые должны поддерживаться во время выполнения. Система управления с обратной связью – это реактивный механизм принятия решений, поскольку она ожидает, пока возмущение не повлияет на выходы системы, чтобы принять необходимые решения.

Другой тип систем управления – это система управления с прямой связью (рассматриваемая как проактивный механизм контроля). Также используется комбинация этих типов, т. е. системы управления с обратной связью и упреждающей связью (которая устраняет ограничения обеих схем) [4, 16]. ТУ используется для анализа многих аспектов сред облачных вычислений (рисунок 2).

Сравнение подходов. Для сравнения рассмотрим возможность учета следующих особенностей облачных систем: проведение оценки производительности многокомпонентных приложений, моделирование рабочей нагрузки, обработки заданий и подготовки ВМ, учет взаимного влияния ВМ на производительность из-за совместного использования физических ресурсов, управление автономными ресурсами – возможность изменения характеристик отдельных ресурсов, консолидация серверов – определение оптимального размещения нескольких ВМ на одном физическом сервере в целях энергосбережения, стратегия масштабирования – вертикальное (замена физических серверов на более мощные, наращивание ресурсов имеющихся серверов) или горизонтальное масштабирование (использование балансировщиков нагрузки, реплик серверов, распределение данных по разным серверам) (таблица 2) [1, 4].

Заключение. Оценка производительности облачных приложений производится как на этапе проектирования, так и во время эксплуатации системы для мониторинга и обеспечения гарантированного качества обслуживания. Аналитические методы оценки производительности являются наименее затратными, так как позволяют провести оценку производительности без создания детализированных сложных имитационных моделей и проведения дорогостоящих измерений на существующих системах. При проведении оценки в облачной среде должны учитываться не только время отклика приложения (среднее или максимальное), модели рабочей нагрузки и дисциплина обслуживания, но и время на подготовку виртуальных машин, их влияние на производительность из-за совместного использования физических ресурсов, консолидация северов, стратегия масштабирования. Рассмотрены методы аналитического моделирования, основанные на теории массового обслуживания, теории управления,

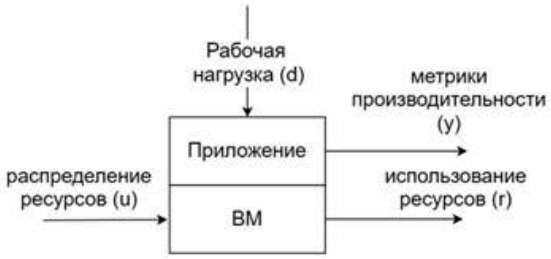

Рисунок 2 – Использование ТУ для автоматизации управления ресурсами и уровнем обслуживания в совместно используемой виртуализированной инфраструктуре для трехуровневого веб-приложения [4]

Таблица 2 – Сравнение методов аналитического моделирования

Возможности моделирования	ТМО	ТМО + ССВ	ТУ	СИ	ИРВ
Многоуровневые веб-приложения	Да	Да	Да	Да	Да
Время отклика – среднее	Да	Да	Нет	Нет	Нет
Время отклика Жесткие/мягкие гарантии	Нет	Нет	Мягкие гарантии	Оба	Оба
Рабочая нагрузка	Искусств.	Искусств.	Реал + Искусств.	Реал + Искусств.	Реал + Искусств.
Обработка	Искусств	Искусств.	Реал + Искусств.	Реал + Искусств.	Реал + Искусств.
Подготовка ВМ	Да	Да	Нет	Нет	Нет
Взаимное влияние ВМ	Да	Да	Да	Да	Да
Автономное управление ресурсами	Да	Да	Да	Да	Да
Консолидация серверов	Да	Да	Да	Да	Да
Стратегия масштабирования	Да	Да	Да	Да	Да

сетевом исчислении, исчислении реального времени. Все эти методы могут применяться для создания моделей приложений и облачной инфраструктуры, но следует отметить, что сетевое исчисление и исчисление реального времени представляются наиболее перспективными для оценки производительности облачных сервисов, так как они дают гарантии жесткого реального времени.

Литература

1. *Qiang Duan*. Cloud service performance evaluation: status, challenges, and opportunities – a survey from the system modeling perspective / Duan Qiang // Digital Communications and Networks. Vol. 3. Iss. 2, May 2017. P. 101–111. URL: https://www.sciencedirect.com/science/article/pii/S2352864816301456 (дата обращения: 5.11.2020).
2. *Ворожцов А.С.* Оценка производительности облачных центров обработки данных / А.С. Ворожцов, Н.В. Тутова, А.В. Тутов // T-Comm. Телекоммуникации и транспорт, 2014. URL: https://cyberleninka.ru/article/v/otsenka-proizvoditelnosti-oblachnyh-tsentrov-obrabotki-dannyh (дата обращения: 5.11.2020).
3. *Li Z.* On a catalogue of metrics for evaluating commercial cloud services. / Z. Li, L. O'Brien, H. Zhang, R. Cai // Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing, 2012. P. 164–173.
4. *Гарай Г.Р.* Сравнительный анализ методов оценки производительности многоуровневых облачных приложений / Г.Р. Гарай, А. Черных, А.Ю. Дроздов // Труды ИСП РАН. 2015. Том 27. Вып. 6. С. 199–224. DOI: 10.15514/ISPRAS-2015-27(6)-14.

5. *Гайдамако В.В.* Инфраструктура Sensor-Cloud – облачные информационно-измерительные системы / В.В. Гайдамако // Проблемы автоматики и управления. 2018. № 2 (35). С. 109–118.
6. *Xiong K.* Service performance and analysis in Cloud computing / K. Xiong, H. Perros // Proceedings of the IEEE 2009 World Congress on Services. 2009. P. 693–700.
7. *Goswami V.* Performance analysis of cloud with queuedependent virtual machines / V. Goswami, S.S. Patra, G.B. Mund // Proceedings of the 1st International Conference on Recent Advances in Information Technology. 2012.
8. *Yang F.* Tan Performance evaluation of Cloud service considering fault recovery / Yang F. Tan, Y.-S. Dai // Supercomput. 2013. 65 (1). 426–444.
9. *Khazaei H.* Performance analysis of Cloud computing centers using M/G/m/m+r queuing systems / H. Khazaei, J. Misic, V.B. Misic // IEEE Trans. Parallel Distrib. Syst. 2012. 23 (5). P. 936–943.
10. *Boudec J.L.* Network Calculus: A Theory of Deterministic Systems for the Internet / J.L. Boudec, P. Thiran // Springer Verlag. Berlin, 2001.
11. *Росляков А.В.* Сетевое исчисление (Network Calculus) и его применение для оценки сетевых характеристик / А.В. Росляков, А.А. Лысиков. Самара: ПГУТИ, 2019. 222 с.
12. *Haid W.* Modular Performance Analysis with Real-Time Calculus / W. Haid, S. Perathoner, N. Stoimenov, L.Thiele // PhD Course on Automated Formal Methods for Embedded Systems DTU. Lyngby, Denmark. June 11, 2007.
13. *Росляков А.В.* Применение теории стохастических сетевых исчислений к анализу характеристик VPN / А.В. Росляков, А.А. Лысиков // T-Comm. 2013. № 7.
14. *Ghosh R.* Modeling and performance analysis of large scale IaaS Clouds, Future Gener / R. Ghosh, F. Longo, V.K. Naik, K.S. Trivedi // Comput. Syst. 2013. 29 (5). P. 1216–1234.
15. T. Lorido-Botran J. Miguel-Alonso and J.A. Lozano, A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments / J. Lorido-Botran, Miguel-Alonso and J.A. Lozano // Journal of Grid Computing. 2014. Vol. 12. P. 559–592.
16. Singh Parminder. Research on Auto-Scaling of Web Applications in Cloud: Survey, Trends and Future Directions / Parminder Singh, Kiran Jyoti, Anand Nayyar // Scalable Computing. 2019. 20(2). P. 399–432.