ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 3(17), 2013, с.??—??

УДК 004.051, 004.453

С.В. Ковальчук, А.М. Чиркин, К.В. Князьков

Моделирование производительности вычислительных сервисов в облачных средах второго поколения

Аннотация. В работе рассматриваются вопросы моделирования производительности вычислительных сервисов, используемых в составе композитных приложений выполняющихся в средах облачных вычислений. Предлагаемый подход основан на использовании комплексных параметрических моделей производительности, позволяющих производить априорную оценку поведения вычислительных сервисов с учетом динамически меняющихся характеристик распределенной вычислительной среды.

Ключевые слова и фразы: облачные вычислений, распределенные вычисления, сервис-ориентированная архитектура, моделирование производительности, композитное приложение.

Введение

Развитие облачных технологий второго поколения, реализующих модель AaaS (Aplication as a Service) выдвигает новые требования к методам оценки характеристик ресурсов, необходимых для получения результата в установленное время с заданным качеством. Они обусловлены тем, что в рамках модели AaaS пользователь работает с композитными приложениями (КП), блоки которых могут выполняться разное время на различных вычислительных ресурсах. Это отличается от традиционных моделей IaaS и PaaS, где требования к вычислительным ресурсам задаются априори пользователем, и от модели SaaS, когда ресурсы согласованы с конкретным программным обеспечением, предоставляемым через облачную среду. Ключевой метрикой для такой оценки является характерное время решения задачи для различных вариантов вы-

боров ресурсов. В силу динамической изменчивости облачной среды время исполнения КП имеет стохастический характер, и должно интерпретироваться в терминах модели случайной величины [1]. Распространенным подходом к решению такой оценки времени выполнения является использование параметрических [2] или стохастических [3] моделей производительности. Тем не менее, сложность и динамичность облачной инфраструктуры, с одной стороны и разнообразие вариантов пользовательских КП, с другой, оставляют эту задачу открытой. В данной работе рассматривается подход, а) обеспечивающий моделирование производительности и прогноз времени выполнения КП; б) допускающий адаптацию моделей к изменяющимся характеристикам вычислительной инфраструктуры; в) предоставляющий возможность оценки неопределенности времени работы КП.

1. Моделирование производительности облачных сервисов и КП

Выделим ряд основных факторов, влияющих на время работы прикладных сервисов и КП в облачной среде (см. таблицу 1). Каждый из этих факторов обладает специфическим характером изменчивости; кроме того, они различаются по величине вклада в общее время работы КП. Однако в общем виде они могут быть представлены как случайные величины, опционально зависящие от параметров четырех категорий: характеристик алгоритма вычислительного пакета (или наборы пакетов, если речь идет о КП), используемого ресурса, входных данных и самой платформы облачных вычислений. Сумма этих величин позволяет произвести оценку времени работы сервиса.

Для облачных сервисов можно выделить две ключевые категории моделей: параметрические – зависящие от входных параметров решаемой задачи (данных) и случайные – характеризующиеся некоторым распределением, инвариантным для конкретного сервиса в определенном состоянии (форма этого распределения меняется в

случае изменения состояния сервиса и его характеристик, но не зависит от входных данных для конкретной задачи). Особым случаем является модель, определяющая время ожидания в очереди, поскольку ее работа зависит от ряда внешних факторов: структуры потока входящих задач, алгоритма планирования и пр. В общем случае очередь можно рассматривать как систему массового обслуживания (СМО) и применять соответствующие подходы для ее моделирования и оценки времени ожидания, учитывая модели отдельных факторов для задач, находящихся в очереди.

Таблица 1. Факторы, влияющие на производительность облачных сервисов и композитных приложений

Фактор	**Зависимость от характеристик**				**Значимость вклада**	**Значимость изменчивости**	**Модель**
	алгоритма	ресурса	входных данных	системы			
Вычисления	+	+	+	-	+++	зависит от алгоритма сервиса	Параметрическая
Передача входных и выходных данных	+	+	+	+	++	++	Параметрическая
Ожидание в очереди	-	-	-	+	++	++	СМО
Доступ к ресурсу	-	+	-	-	+	+	Случайная
Накладные расходы	-	+	-	+	+	+	Случайная

В данной работе рассматривается унифицированный подход, который позволяет получать оценку времени выполнения облачных сервисов и КП, и обеспечивает поддержку эксплуатации облачной среды в части оценки и планирования стоимости использования ресурсов. При этом сложность данной задачи обусловлена не только изменчивостью характеристик и состояния ресурсов, но и неоп-

ределенностью, связанной с представлением КП в форме абстрактных потоков работ (workflow, WF); при этом ресурсы выделяются непосредственно в ходе исполнения самого приложения. Для его реализации использована концепция Intelligent PSE (iPSE) [4] и реализованная на ее основе технологическая платформа облачных вычислений CLAVIRE [5]. Концепция iPSE использует формальные знания для описания специфики работы прикладных сервисов в распределенной среде. К ним относятся и параметрические модели в составе базы знаний платформы CLAVIRE, которые формализуются при описании прикладных пакетов на специализированном предметно-ориентированном языке EasyPackage и используют его возможности для определения высокоуровневых параметров, явно задаваемых в композитном приложении или определяемых по входным данным. На рис. 1 приведена общая архитектура подсистемы моделирования производительности и оценки времени выполнения задач, реализованная в CLAVIRE.

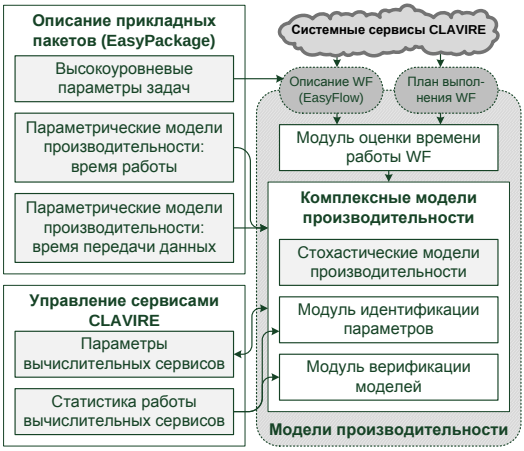

Рис. 1. Программная система для оценки производительности облачной инфраструктуры на базе платформы CLAVIRE

В соответствии с рис. 1, параметры вычислительного пакета могут быть определены экспериментально (на этапе его встраива-

ния в облачную среду) и зафиксированы в описании на языке EasyPackage. Однако параметры ресурсов, а также параметры самой управляющей инфраструктуры требуют идентификации, которая должна происходить в автоматическом режиме в целях адаптации к меняющимся характеристикам облачной среды, которая может выполняться на основе обновляемых данных – статистики использования прикладных сервисов CLAVIRE.

2. Идентификация и анализ моделей производительности

В соответствии с табл. 1, основой для определения времени выполнения облачного сервиса является параметрическая модель производительности. Она может быть представлена в явной форме (в виде выражения, связывающего время выполнения с характеристиками входных данных и параметрами инфраструктуры), и в неявной форме (в виде некоторого решающего алгоритма).

2.1. Модели производительности в явной форме

Модель в явной форме представляет собой выражение, составляемое при описании прикладного пакета в CLAVIRE. Случайная величина $T_{\mu,\sigma}$ (время выполнения) может быть характеризуема функциями $\mu(\Xi_D,\Xi_S)$ и $\sigma(\Xi_D,\Xi_S)$, определяющими, соответственно, среднее значение и СКО. Структура этих функций определяется через выражение для сложности алгоритма. Они отражают зависимость от наборов параметров, связанных с входными параметрами (Ξ_D), и характеристиками вызываемого вычислительного сервиса (Ξ_S). Задача идентификации параметров Ξ_S, решается на основе анализа статистики запусков, содержащих значение параметров Ξ_D и измеренное время выполнения. Методы решения задачи идентификации изложены в [6].

2.2. Модели производительности в неявной форме

Построение параметрической модели производительности требует знаний об алгоритме, реализуемом в вычислительном пакете. Если такие сведения отсутствуют, можно воспользоваться методами машинного обучения для построения модели в неявной форме. Такие модели до обучения не содержат никакой информации о зависимости оцениваемой величины от параметров и аргументов. В ходе обучения модель формирует зависимости от аргументов Ξ_D, но параметры Ξ_S у нее отсутствуют (они учитываются как *внутреннее состояние* модели). В частности, для построения моделей производительности допустимо применять метод случайного леса (random forest) [7]. Основная идея этого метода заключается в построении большого количества деревьев принятия решений и композиции их решений. Достоинство этого метода в том, что он прост в настройке (его не нужно настраивать под каждую модель вручную) и быстро обучаем.

2.3. Валидация моделей

Необходимость валидации (проверки состояния) модели возникает в связи с возможными изменениями характеристик облачной инфраструктуры. Они могут быть обусловлены как эволюционными факторами (вариация загрузки сетей, связанная с суточной ритмикой), так и кардинальными факторами (изменение настроек вычислительных ресурсов). Потому необходимо обновить (откорректировать) параметры Ξ_S модели или ее внутреннее состояние. Случайную величину $T_{\mu,\sigma}$ можно представить в форме

$$T_{\mu,\sigma}(\Xi_t) = \mu(\Xi_t) + \sigma(\Xi_t) \cdot \xi(t), \qquad (1)$$

где Ξ_t – параметры, связанные с запуском, происходившим в момент времени t, а $\xi(t)$ – относительная величина ошибки, которую можно представить как процесс во времени, явно не зависящий от параметров модели. Цель валидации модели – определить

по текущим данным и параметрам модели, нужно ли переопределять коэффициенты модели в связи с изменением поведения $T_{\mu,\sigma}$. Для этого используются статистические критерии на основе усеченной выборки за определенное время, в течение которого случайный процесс $\xi(t)$ считается стационарным, в частности, ранговый U-критерий Манна-Уитни и критерий согласия Колмогорова.

2.4. Коррекция коэффициентов моделей

Обновление модели целиком на основе выборки данных может стать относительно трудоемкой задачей ввиду ее неоднородности, а критерии для валидации модели подразумевают некоторые интервалы, в которых значения модели могут меняться. Потому в ряде случаев целесообразно использовать дополнительно линейные фильтры для коррекции модели, в частности, экспоненциально затухающее скользящее среднее.

3. Экспериментальные исследования моделей производительности

На рис. 2(а) изображены результаты сопоставления временных характеристик, полученных по результатам измерений и по моделям производительности. В качестве объекта эксперимента использовался пакет ComsolFloodSimulator, позволяющий моделировать затопление территорий при возникновении наводнений [8], доступный в форме облачного сервиса платформы CLAVIRE.

Время работы данного пакета зависит линейно от времени моделирования, в то время как остальные параметры (при заданной геометрии расчетной области) не оказывают существенного влияния на время работы. Для сопоставления рассмотрены две модели – параметрическая линейная модель производительности, зависящая от одного параметра (в явной форме) и модель в неявной форме на основе метода случайного леса. На рис. 2а проиллюстрировано соответствие времени, ожидаемого в соответствии с результатами моделирования, и измеренного времени вычислений. Ошибка измеряется как относительное среднее отклонение от измеренно-

го значения времени. Проведенные экспериментальные исследования показывают, что в целом точность формализованной модели больше (ошибка 3,7% против 9%), но эффективность второй модели также приемлема. Это позволяет рассматривать модели такого класса как альтернативу в случае отсутствия сведений об алгоритме, которые нужны для построения моделей в явной форме.

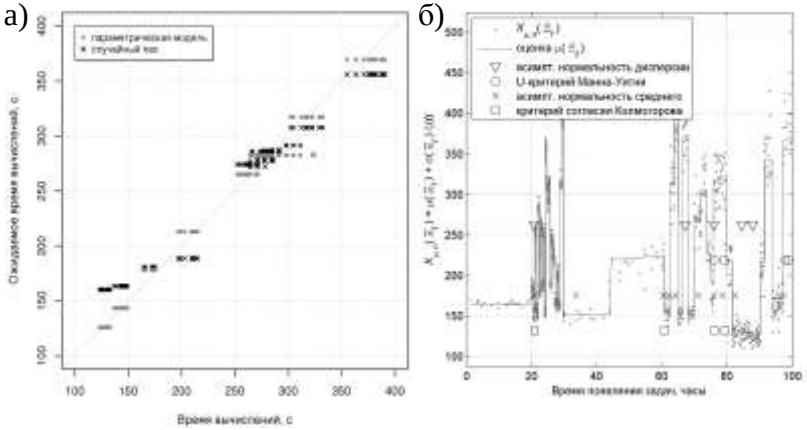

Рис. 2. Исследования параметрических моделей производительности: а) время работы сервиса ComsolFloodSimulator; б) коррекция параметров в процессе работы системы

На рис. 2(б) проиллюстрирован процесс валидации модели. Специальными символами (∇, $\bigcirc$, $\times$, $\square$) на графике отмечены моменты времени срабатывания различных критериев, точками и линией – моделируемая величина (время вычисления) и ее оценка соответственно. На оси абсцисс отмечено время запусков пакета. Как видно из рисунка, время и частота срабатывания критериев заметно отличаются. Критерии используются для определения необходимости повторной идентификации модели производительности. Очевидно, что в целях повышения производительности системы стоит выбирать вариант с достаточно редкой повторной иден-

тификацией, и тем не менее, обеспечивающий точность работы модели. Как видно из графика, выбранные критерии (U-критерий Манна-Уитни и критерий согласия Колмогорова) обеспечивают достаточно редкий вызов процедуры повторной идентификации по сравнению с более примитивными критериями. Кроме того, экспериментальные исследования процедуры валидации показали, что, варьируя уровень доверия критериев или размер активной выборки, можно изменить точность прогнозирования и количество срабатываний критериев, что в целом позволяет снизить частоту обновлений модели.

Заключение

Предложенное решение, будучи интегрированным в платформу вычислений CLAVIRE, позволяет производить априорную оценку поведения (в первую очередь, времени работы) облачных сервисов и КП, при этом адаптируясь к изменяющимся характеристикам распределенной вычислительной среды. В качестве альтернативных решений используются как модели, автоматически построенные на основе истории запусков с помощью алгоритмов машинного обучения, так и модели в явном виде формализованные при описании прикладных пакетов, доступных в облачной среде. Как результат, важной особенностью данного решения является возможность динамического уточнения времени работы КП по мере накопления статистики их использования.

Работа выполнена в рамках реализации постановления №220 Правительства РФ (договор №11.G34.31.0019) при поддержке ФЦП «Научные и научно-педагогические кадры инновационной России" на 2009 - 2013 годы», соглашение 14.B37.21.0596 от 17.08.2012 г.

Список литературы

[1] Чуров Т.Н. [и др.] *Особенности вероятностного анализа производительности и надежности проблемно-ориентированных сред облачных вычислений* // Известия высших учебных заведений. Приборостроение. 2011. Т. 54. № 10. С. 51-57.

[2] Kishimoto Y., Ichikawa S. *Optimizing the Configuration of a Heterogeneous Cluster with Multiprocessing and Execution-Time Estimation* // Parallel Computing. 2005. Vol. 31. No. 7. pp. 691-710.

[3] Trebon N.D. *Enabling Urgent Computing within the Existing Distributed Computing Infrastructure*, Ph.D. Dissertation, University of Chicago. August 2011. 117 p.

[4] Бухановский А.В., Ковальчук С.В., Марьин С.В. *Интеллектуальные высокопроизводительные программные комплексы моделирования сложных систем: концепция, архитектура и примеры реализации* // Известия высших учебных заведений. Приборостроение. 2009. №10. С. 5-24.

[5] Васильев В.Н. [и др.] *CLAVIRE: облачная платформа для обработки данных больших объемов* // Информационно-измерительные и управляющие системы. 2012. Т. 10. №11. С. 7-16.

[6] Kovalchuk S.V. [et al.] *Deadline-Driven Resource Management within Urgent Computing Cyberinfrastructure* // Procedia Computer Science. Vol. 18. Proceedings of the International Conference on Computational Science. ICCS 2013. 2013. pp. 2203-2212.

[7] Breiman L. *Random Forests* // Machine Learning. 2001. Vol. 45. pp. 5-32.

[8] Krzhizhanovskaya V.V. [et al.] *Distributed Simulation of City Inundation by Coupled Surface and Subsurface Porous Flow for Urban Flood Decision Support System* // Procedia Computer Science, Vol. 18, Proceedings of the International Conference on Computational Science. ICCS 2013. 2013. pp. 1046–1056.

Об авторах:

Ковальчук Сергей Валерьевич

Национальный исследовательский университет информационных технологий, механики и оптики, с.н.с., к.т.н.

e-mail: kovalchuk@mail.ifmo.ru

Чиркин Артём Михайлович

Национальный исследовательский университет информационных технологий, механики и оптики, студент

e-mail: chirkin.art@gmail.com

Князьков Константин Валерьевич

Национальный исследовательский университет информационных технологий, механики и оптики, с.н.с., к.т.н.

e-mail: constantinvk@gmail.com

S.V. Kovalchuk, A.M. Chirkin, K.V. Knyazkov. Modeling of Computational Services Performance within Second-Generation Cloud Computing Environments.

Abstract. The issues of performance modeling and simulation for computational services within composite applications executed in cloud computing environments are considered. The proposed approach is based on the usage of complex parametric models, which allow to estimate behavior of computational services taking into account dynamically changing characteristics of the distributed environment.

Key Words and Phrases: cloud computing, distributed computing, service-oriented architecture, performance modeling and simulation, composite application.