МНОГОКРИТЕРИАЛЬНАЯ ОПТИМИЗАЦИЯ РАЗМЕЩЕНИЯ ВИРТУАЛЬНЫХ МАШИН ПО ФИЗИЧЕСКИМ СЕРВЕРАМ В ОБЛАЧНЫХ ЦЕНТРАХ ОБРАБОТКИ ДАННЫХ

Тутов Андрей Владимирович,
Московский технический университет связи и информатики, Москва, Россия, ***andrew_vidnoe@mail.ru***

Тутова Наталья Владимировна,
Московский технический университет связи и информатики, Москва, Россия, ***e-natasha@mail.ru***

Ворожцов Анатолий Сергеевич,
Московский технический университет связи и информатики, Москва, Россия, ***as.vorojcov@mail.ru***

Андреев Илья Александрович,
Московский технический университет связи и информатики, Москва, Россия, ***1c@mtuci.ru***

DOI: 10.36724/2072-8735-2021-15-1-28-34

Manuscript received 02 September 2020;
Revised 05 October 2020;
Accepted 07 December 2020

Ключевые слова: *виртуализация, виртуальные машины, задача о назначениях, центры обработки данных, оптимальное размещение, многокритериальная оптимизация*

Рассмотрена задача размещения виртуальных машин на физических серверах в системе управления ресурсами облачного центра обработки данных. Система имеет двухуровневую архитектуру, состоящую из глобального и локальных контроллеров. Локальные контроллеры анализируют состояние физических серверов, на которых они расположены и определяют возможные состояния недогрузки, перегрузки и перегрева на основании прогноза для следующего окна наблюдения. В случае выявления одного из перечисленных состояний, локальный контроллер сообщает об этом глобальному контроллеру, который выбирает серверы назначения, на который будет производиться размещение виртуальных машин посредством миграции. Размещение виртуальных машин предложено проводить по таким критериям качества облачных сервисов, как минимум остатка неиспользуемых ресурсов и нарушений SLA-соглашений. Сформулирована математическая постановка задачи оптимизации, которая по структуре, необходимым условиям, характеру переменных эквивалентна известной основной задаче о назначениях. Сведение задачи о назначениях к замкнутой транспортной задаче позволило в условиях жесткого реального времени решить при многих критериях задачу размещения виртуальных машин и значительно увеличить ее размерность в сравнении с эвристическими алгоритмами, что дает возможность поддерживать качество современных облачных сервисов в условиях стремительного роста физических и виртуальных ресурсов центров обработки данных. Разработанная математическая постановка задачи, результаты вычислительных экспериментов могут быть включены в математическое обеспечение процессов живой миграции виртуальных машин.

Информация об авторах:

Тутов Андрей Владимирович, *старший преподаватель кафедры "Корпоративные информационные системы", Московский технический университет связи и информатики, Москва, Россия*

Тутова Наталья Владимировна, *доцент кафедры "Корпоративные информационные системы", доцент, к.т.н., Московский технический университет связи и информатики,Москва, Россия*

Ворожцов Анатолий Сергеевич, *доцент кафедры "Интеллектуальные системы в управлении и автоматизации", доцент, к.т.н.,Московский технический университет связи и информатики, Москва, Россия*

Андреев Илья Александрович, *и.о. зав. кафедрой "Корпоративные информационные системы", к.э.н., Московский технический университет связи и информатики, Москва, Россия*

Для цитирования:
Тутов А.В., Тутова Н.В., Ворожцов А.С., Андреев И.А. Многокритериальная оптимизация размещения виртуальных машин по физическим серверам в облачных центрах обработки данных // T-Comm: Телекоммуникации и транспорт. 2021. Том 15. №1. С. 28-34.

For citation:
Toutov A.V., Toutova N.V., Vorozhtsov A.S., Andreev I.A. (2021) Multi-objective optimization of virtual machine placement on physical servers in cloud data centers. *T-Comm*, vol. 15, no.1, pp. 28-34. (*in Russian*)

Введение

Центры обработки данных (ЦОД) должны предоставлять достаточное количество ресурсов для бесперебойной работы размещенных в них приложений, нагрузка на которых может значительно меняться во времени. В случаях, когда эти изменения выходят за рамки допустимых значений работоспособность приложений снижается, что требует динамического перераспределения ресурсов. В облачных ЦОД распределение ресурсов осуществляется путем миграции виртуальных машин (ВМ) – перемещения виртуальных машин между физическими серверами. Если миграция происходит без отключения и потери доступности ВМ, то такая миграция носит название «живой». Такая миграция позволяет выполнять соглашения об уровне сервиса (SLA), сбалансировать нагрузку между физическими серверами ЦОД, а также размещать виртуальные машины на меньшем числе физических серверов.

Кроме этого, появились сети центров обработки данных и информационные ресурсы могут перераспределяться между ЦОД в соответствии со спросом на ресурсы, например, с учётом часовых поясов. Это заставляет искать эффективные и быстрые алгоритмы распределения ресурсов с учетом растущих размерностей задач.

Процесс динамического распределения ресурсов включает в себя три этапа: мониторинг серверов и обнаружение критических ситуаций, выбор виртуальных машин и выбор серверов назначения [1].

В данной работе основное внимание уделяется третьему этапу выбора серверов для размещения виртуальных машин. Поставлена задача многокритериальной оптимизации, выбран метод ее решения.

Обзор работ

Проблеме оптимизации вычислительных ресурсов центров обработки данных, в частности оптимальному размещению виртуальных машин по физическим серверам, посвящено большое число работ. В связи с ростом Интернет-трафика, появлением «больших данные», развитием и распространение облачных сервисов и систем искусственного интеллекта предъявляются все большие требования к инфраструктуре ЦОД и оптимизации использования ресурсов.

Рассматриваемая задача размещения виртуальных машин по физическим серверам в литературе представлена в двух аспектах: первоначальное или статическое размещение [2-6] и динамическое размещение [7-14]. Данные этапы включены в основной цикл работ по оптимизации ресурсов облачных ЦОД с учетом процесса живой миграции [15].

В большинстве работ, посвященных статическому размещению, используются формулировки задачи об упаковке в контейнеры или о рюкзаке [2-5]. Данные задачи относятся к классу комбинаторных задач оптимизации с бинарными переменными, точное решение которых нельзя получить за полиномиальное время. Поэтому для их решения используются приближенные алгоритмы.

На практике для выбора серверов при динамическом размещении виртуальных машин получили распространение эвристические алгоритмы, такие как FFD (First Fit Decreasing – первый подходящий по убыванию) и BFD (Best Fit Decreasing – наилучший подходящий по убыванию), а также их модификации [7, 9, 12].

Время решения задачи оптимизации с учетом процессов миграции является одним из основных факторов, влияющих на качество принятия решений в реальном времени. За один цикл работы глобального контроллера, составляющий от двух до пяти минут необходимо обнаружить проблему на серверах (перегрузка, недогрузка или перегрев), выбрать ВМ для миграции и серверы для размещения ВМ, а также переместить виртуальные. Задержки в принятии решении могут привести к штрафным санкциям за нарушение SLA-соглашений и дополнительным затратам на охлаждение серверов.

Многокритериальные задачи размещения рассматривались в работах [2,3,5,6,8]. Для решения подобных многокритериальных задач чаще всего используются методы формирования обобщенного критерия [2,3,5,8]. В работе [6] для выбора сервера использовался метод анализа иерархий, что ограничивает применение в реальном времени.

Однако в приведенных работах на этапе динамического размещения слабо освящены вопросы размерности и времени решения задач. Поэтому целесообразно поставить такую задачу, которая позволяла бы масштабировать ресурсы в широких пределах и давать возможность получить точное решение в реальном времени.

Архитектура системы

Динамическое размещение вычислительных ресурсов с включением процесса живой миграции виртуальных машин является основным этапов цикла системы управления облачным центром обработки данных, структурная схема которой приведена на рисунке 1.

Типовая система управления ресурсами облачного ЦОД имеет двухуровневую архитектуру, состоящую из глобального и локальных контроллеров. На локальных контроллерах постоянно анализируются данные системы мониторинга о состоянии физических серверов, на которых они расположены. Определяются возможные состояния недогрузки, перегрузки и перегрева на основании прогноза для следующего окна наблюдения. Проверка показателей системы осуществляется последовательно в соответствии с важностью критериев (перегрев, перегрузка, недогрузка сервера), при этом процесс наблюдения осуществляется непрерывно, в том числе в случае выполнения действий по перемещению ВМ (рис. 2).

В случае выявления одного из перечисленных состояний, локальный контроллер посылает сообщение глобальному контроллеру, на котором инициируется процесс миграции виртуальных машин: выбираются виртуальные машины для миграции и серверы назначения. Выбор сервера назначения является ответственным этапом, поскольку неудачный выбор сервера может привести к новым нежелательным миграциям, поскольку сами миграции дополнительно нагружают систему и приводят к ухудшению производительности и простою виртуальных машин.

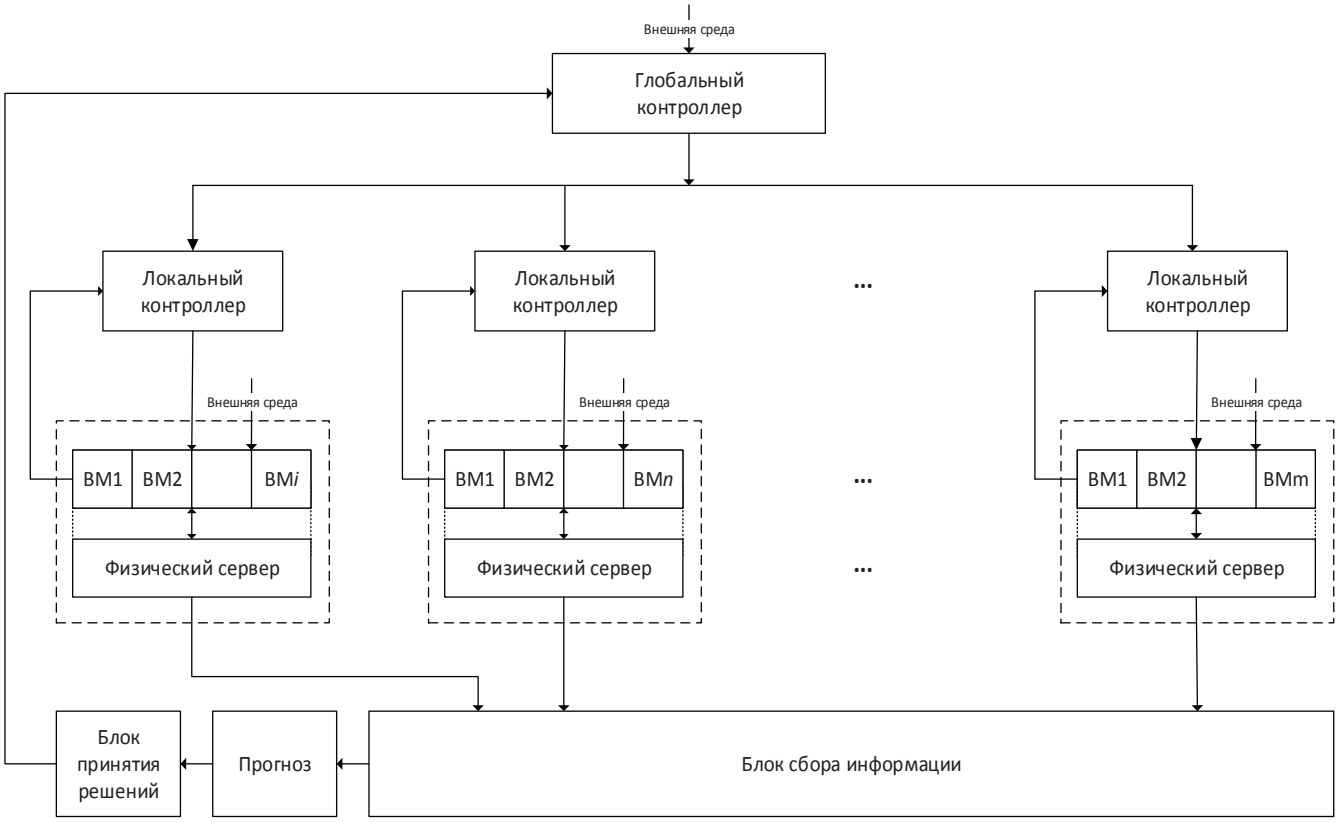

Рис. 1. Двухуровневая система управления ресурсами облачного ЦОД

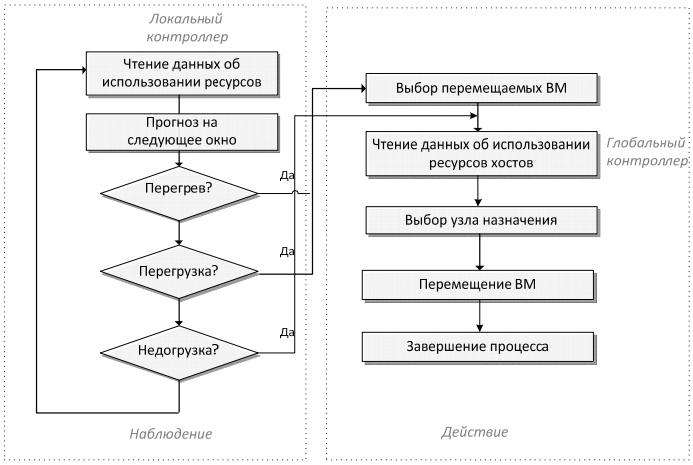

Рис. 2

Постановка задачи размещения виртуальных машин

Пусть имеется N активных физических серверов и такое же количество виртуальных машин для размещения путем миграции. Данные серверы являются работающими и могут обслуживать другие виртуальные машины.

Ресурсы i-ой, i=1,…,N, виртуальной машины зададим производительностью её процессора VM_i^{CPU} и объемом оперативной памяти VM_i^{RAM}. Такие же характеристики будем считать заданными для каждого сервера $PM_j^{CPU\_0}$ и $PM_j^{RAM\_0}$. Предположим, что в рассматриваемый момент времени у каждого сервера задействована часть ресурсов другими виртуальными машинами, которую обозначим для процессора $PM_j^{CPU\_1}$ и $PM_j^{RAM\_1}$ для памяти. Предположим также, что свободной части ресурсов достаточно для размещения любой ВМ из N имеющихся.

Учтем цикличность процесса динамического размещения виртуальных машин, что дает возможность предположить: за один цикл работы контроллера на i-ый сервер одновременно может быть размещено не более одной ВМ, и i-ая ВМ может быть размещена только на одном, а не на двух и более серверах (рис. 3).

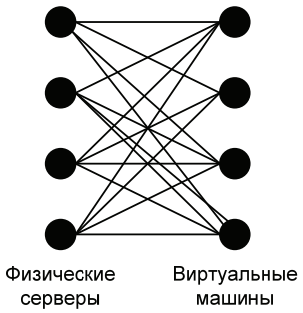

Рис. 3. Иллюстрация задачи размещения виртуальных машин по серверам

Требуется найти оптимальное размещение виртуальных машин с учетом минимального времени принятия решения по процессу размещения, которое не превышает цикл работы контроллера 2-5 минут [8, 12].

Выбор критериев оптимизации

Процесс размещения виртуальных машин на физических серверах описывается рядом показателей: эффективность использования ресурсов [2,3,5], энергопотребление [2, 3, 5, 9-14], равномерное распределение температуры [3,8], выполнение SLA-соглашений [2,3], балансировка нагрузки [7], минимизация трафика [14] и другие. Поэтому задача о размещении должна решаться по двум и более критериям. Наибольшее распространение получили критерии эффективности использования ресурсов и нарушения SLA-соглашений.

Предположим, что размещаемая на сервере виртуальная машина занимает всю выделенную ей память и процессорное время. Тогда обозначим u_{ij}^{CPU} загрузку процессора сервера j после размещения виртуальной машины i, а u_{ij}^{RAM} загрузку памяти сервера j после размещения виртуальной машины i.

$$u_{ij}^{CPU} = \frac{PM_j^{CPU\_1} + VM_i^{CPU}}{PM_j^{CPU\_0}}$$

$$u_{ij}^{RAM} = \frac{PM_j^{RAM\_1} + VM_i^{RAM}}{PM_j^{RAM\_0}}$$

Критерий неэффективности использования ресурсов отражает насколько несбалансированно используются ресурсы на каждом сервере. Остаток ресурсов на каждом сервере должен быть сбалансирован по нескольким видам ресурсов. Например, неудачным размещением виртуальных машин является такое размещение, при котором вся память сервера использована, а процессор малозагружен.

Критерий неэффективности использования ресурсов для сервера можно сформулировать следующим образом:

$$f_{res}^{j}(u_{ij}^{CPU}, u_{ij}^{RAM}) = 1 - u_{ij}^{CPU} \cdot u_{ij}^{RAM}$$

Данный критерий отражает насколько полно загружены ресурсы серверов различных типов. Значения данного критерия лежат в диапазоне от 0 до 1. Чем ближе значение критерия к нулю, тем лучше загружены ресурсы сервера.

Наиболее часто требование на качество обслуживания облачного сервиса задаётся в виде среднего времени отклика, которое зависит от загрузки процессора. При превышении некоторого порога загрузки происходит снижение производительности приложения и увеличение времени отклика. В литературе используются пороговые значения загрузки в диапазоне 80-90% [8, 12]. В качестве критерия **нарушения SLA-соглашений** выбрана следующая логистическая функция.

$$f_{SLA}^{j}(u_{ij}^{CPU}) = 1 - \frac{1}{1 + e^{-(u_{ij}^{CPU} - 0{,}8)}},$$

Значение этой функции также лежат в диапазоне от 0 до 1, в точке порогового значения $u_{ij}^{CPU} = 0{,}8$ значение функции равно 0,5 и быстро увеличивается при превышении порогового значения. Данный критерий должен быть минимизирован.

Математическая постановка задачи

Введем переменные задачи x_{ij}, которые будут соответствовать назначению ВМ на сервер. В этом случае $x_{ij} = 1$, если i-ая ВМ назначается на выполнение j-му серверу, и $x_{ij} = 0$, если i-ая ВМ не назначается на выполнение j-му серверу.

Поставленная задача по структуре, необходимым условиям, характеру переменных эквивалентна известной основной задаче о назначениях, математическая постановка которой для представленной предметной области имеет вид:

$$\sum_{i=1}^{N}\sum_{j=1}^{N} f_{res}^{j}(u_{ij}^{CPU}, u_{ij}^{RAM}) \cdot x_{ij} \to \min_{x \in \Delta_\beta},$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} f_{SLA}^{j}(u_{ij}^{CPU}) \cdot x_{ij} \to \min_{x \in \Delta_\beta},$$

где множество допустимых альтернатив Δ_β формируется следующей системой ограничений:

(1)

$$\begin{cases} \sum_{j=1}^{N} x_{ij} = 1, \forall i \in \{1,2,\ldots,N\} \\ \sum_{i=1}^{N} x_{ij} = 1, \forall j \in \{1,2,\ldots,N\} \\ x_{ij} \in \{0,1\}, \forall i,j \in \{1,2,\ldots,N\} \end{cases} \quad (2)$$

Критерии задачи относится к классу линейных. Коэффициенты критериальной функции являются матрицами размером $N \times N$, их расчет осуществляется по формулам (1-4). Умножение коэффициентов на матрицу размещения X влияет в конечном итоге на значения критериев.

(3)

Метод решения

Известно [16], что задача о назначении может быть сведена к замкнутой транспортной задаче путем замены ограничения $x_{ij} \in \{0,1\}$ на $x_{ij} \geq 0$. Это даёт возможность использовать для получения оптимального решения известное и доступное программное обеспечение. Выберем такой метод решения задачи размещения виртуальных машин по физическим серверам.

Сведение задачи к транспортной позволяет также реализовать её многокритериальность, путем использования такого подхода как свёртка критериев со всеми её преимуществами [17].

В итоге постановка задачи будет выглядеть следующим образом:

$$\alpha_1 \sum_{i=1}^{N}\sum_{j=1}^{N} f_{res}^{j}(u_{ij}^{CPU}, u_{ij}^{RAM}) \cdot x_{ij} + \alpha_2 \sum_{i=1}^{N}\sum_{j=1}^{N} f_{SLA}^{j}(u_{ij}^{CPU}) \cdot x_{ij} \to \min_{x \in \Delta_\beta} \quad (4)$$

при ограничениях

$$\begin{cases} \sum_{j=1}^{N} x_{ij} = 1, \forall i \in \{1,2,\ldots,N\} \\ \sum_{i=1}^{N} x_{ij} = 1, \forall j \in \{1,2,\ldots,N\} \\ x_{ij} \geq 0, \forall i, j \in \{1,2,\ldots,N\} \end{cases},$$

где α_1, α_2 – веса критериев; $\alpha_1 + \alpha_2 = 1, \alpha_1, \alpha_2 \geq 0$,

Вычислительные эксперименты

Проведено две серии экспериментов для оценки предложенного многокритериального подхода к размещению ВМ с точки зрения эффективности размещения и масштабируемости. Параметры виртуальных машин генерировались случайным образом. Производительность процессора виртуальных машин в ГГц равномерно распределена из следующего набора значений {0.25, 0.5, 1, 1.5, 2, 2.5, 3, 4}, а память из набора значений {0.25, 0.5, 1, 1.5, 2, 2.5, 3, 4}. Число доступных серверов и ВМ задается в качестве начальных значений для имитации ЦОД различных размеров.

В таблице 1 приведены настройки параметров для двух наборов экспериментов. Для каждой серии экспериментов сгенерированы наборы случайных входных данных. Каждый эксперимент был проведен 20 раз. Полученные результаты усреднены.

Таблица 1

Серия экспериментов	Размер ЦОД
1. Эффективность размещения	50 ВМ, 50 серверов
2. Масштабируемость	50~250 ВМ и серверов

В каждом наборе экспериментов было проведено сравнение предложенного многокритериального алгоритма с применяющимися на практике эвристическими алгоритмами задачи упаковки в контейнеры.

Алгоритмы Best Fit Decreasing (BFD) – Список серверов сортируются по убыванию в соответствии с производительностью процессора (bfd_cpu) или памяти (bfd_mem), далее каждая виртуальная машина закрепляется за сервером таким образом, чтобы остаток используемого ресурса (процессора или памяти) был минимальным.

Алгоритмы First Fit Decreasing (FFD) – список серверов сортируются по убыванию в соответствии с производительностью процессора (ffd_cpu) или памяти (ffd_mem), далее каждая виртуальная машина закрепляется за первым походящим сервером в списке.

При решении задачи оптимизации методом свертки, значения вектора весовых коэффициентов изменялись в диапазоне (0,1; 0,9) ~ (0,9; 0,1) с шагом 0,2.

Также вычислялись минимальные и максимальные значения по каждому критерию, относительно которых вычислялись нормализованные значения критериев. Результаты по критериям нарушения SLA-соглашений и эффективности использования ресурсов для алгоритмов bfd_cpu, bfd_mem, ffd_cpu, ffd_mem и $\alpha_1, \alpha_2 = 0{,}1; 0{,}3; 0{,}5; 0{,}7; 0{,}9$ приведены на рис. 4 и 5.

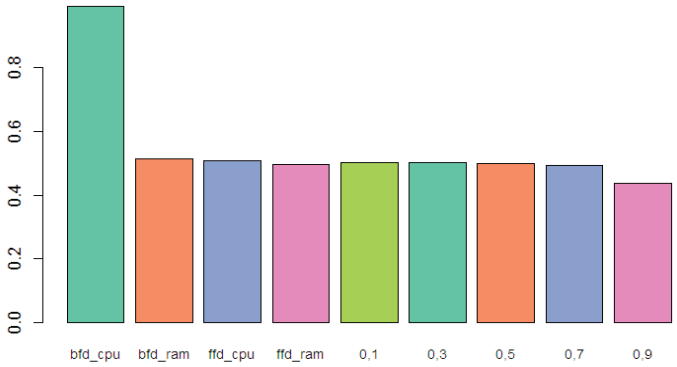

Рис. 4. Нормализованные значения критерия нарушения SLA-соглашений для различных алгоритмов

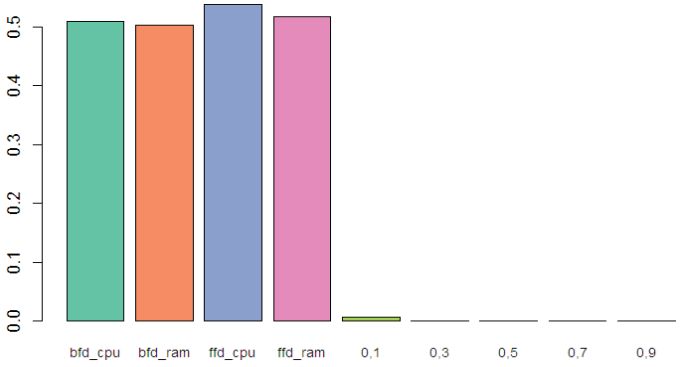

Рис. 5. Нормализованные значения критерия неэффективности использования ресурсов

Как видно, в соответствии с алгоритмом bfd_cpu виртуальные машины размещались плотнее на серверы и как следствие имеется самый высокий уровень нарушения SLA-соглашений. Также алгоритмы bfd_cpu, bfd_mem, ffd_cpu, ffd_mem, являясь однокритериальными, демонстрируют неэффективную загрузку нескольких видов ресурсов (рис. 5).

В этом отношении свертка является предпочтительным методом, дающим эффективное решение по двум критериям.

На основании полученных результатов вычислительных экспериментов, можно сделать вывод, что решение поставленной задачи оптимизации размещения виртуальных машин как транспортную задачу позволяет найти оптимальное значение по двум критериям по сравнению с эвристическими алгоритмами.

Вторая серия экспериментов была направлена на определение времени решения в зависимости от размерности задачи. Транспортная задача решалась симплекс-методом, реализованным в пакете lpSolve для языка R на компьютере Pentium(R) Dual-Core CPU E5700 3 ГГц 4 Гб ОЗУ. Время решения в зависимости от размерности задачи приведено в таблице 2.

Таблица 2

Зависимость времени решения от размерности задачи

Размерность	Время (с)
50 ВМ, 50 серверов	0,73
100 ВМ, 100 серверов	1,72
150 ВМ, 150 серверов	4,84
200 ВМ, 200 серверов	11,44
250 ВМ, 250 серверов	22,17

Зависимость времени решения задачи от размерности является квадратичной $O(n^2)$ с величиной достоверности аппроксимации $R^2 = 0{,}9916$ и приведена на рис. 6. Вычислительная сложность алгоритма решения данной задачи является оптимальным относительно порядка сложности [18].

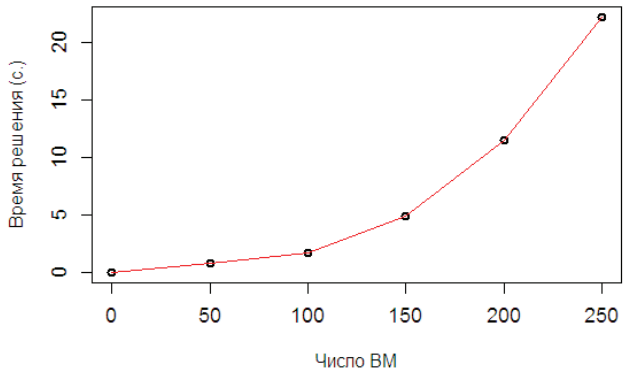

Рис. 6. Зависимость времени решения задачи от размерности

Выводы

Сведение основной задачи о назначении к замкнутой транспортной задаче позволило в условиях жесткого реального времени решить при многих критериях задачу размещения виртуальных машин и значительно увеличить ее размерность в сравнении с эвристическими алгоритмами, что дает возможность поддерживать качество современных облачных сервисов в условиях стремительного роста физических и виртуальных ресурсов центров обработки данных.

Разработанная математическая постановка задачи, результаты вычислительных экспериментов могут быть включены в математическое обеспечение процессов живой миграции виртуальных машин.

Литература

1. *Ворожцов А.С., Тутова Н.В., Тутов А.В.* Динамическое распределение вычислительных ресурсов центров обработки данных // T-Comm: Телекоммуникации и Транспорт. 2016. Т. 10. №.7.

2. *J. Xu and J. Fortes*, “Multi-objective Virtual Machine Placement in Virtualized Data Center Environments”, Proceedings of the 2010 IEEE/ACM Conference on Green Computing and Communications, 179-188.

3. *Ворожцов А.С., Тутова Н.В., Тутов А.В.* Оптимизация размещения облачных серверов в центрах обработки данных // T-Comm: Телекоммуникации и транспорт. №6. 2015.

4. *Camati R. S., Calsavara A., Lima Jr L.* Solving the virtual machine placement problem as a multiple multidimensional knapsack problem // ICN 2014. 2014. С. 264.

5. *Ferdaus M.H.* et al. Virtual machine consolidation in cloud data centers using ACO metaheuristic // European conference on parallel processing. Springer, Cham, 2014. С. 306-317.

6. *Микрюков А.А., Хантимиров Р.И.* Задача первоначального выделения ресурсов в облачных вычислительных средах на основе метода анализа иерархий // Статистика и Экономика. 2015. №.4. С. 184-187.

7. *Gulati A.* et al. Vmware distributed resource management: Design, implementation, and lessons learned // VMware Technical Journal. 2012. Т.1. №.1. С. 45-64.

8. *Xu J., Fortes J.* A multi-objective approach to virtual machine management in datacenters // Proceedings of the 8th ACM international conference on Autonomic computing. 2011. С. 225-234.

9. *Moges F., Abebe S.* Energy-aware VM placement algorithms for the OpenStack Neat consolidation framework // J Cloud Comp 8, 2, 2019.

10. *Shaw R., Howley E., Barrett E.* An energy efficient anti-correlated virtual machine placement algorithm using resource usage predictions // Simulation Modelling Practice and Theory. 2019. Т. 93. С. 322-342.

11. *Alharbi F.* et al. An ant colony system for energy-efficient dynamic virtual machine placement in data centers // Expert Systems with Applications. 2019. Т. 120. С. 228-238.

12. *Beloglazov A., Abawajy J., Buyya R.* Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing //Future generation computer systems. 2012. Т. 28. №.5. С. 755-768.

13. *Murtazaev A., Oh S. Sercon.* Server consolidation algorithm using live migration of virtual machines for green computing // IETE Technical Review. 2011. Т.28. №.3. С. 212-231.

14. *Farzai S., Shirvani M. H., Rabbani M.* Multi-objective communication-aware optimization for virtual machine placement in cloud datacenters //Sustainable Computing: Informatics and Systems. 2020. С. 100374.

15. *Тутов А.В.* Модели и методы распределения ресурсов инфокоммуникационной системы облачных центров обработки данных // Наукоемкие технологии в космических исследованиях Земли. 2018. Т.10. №.6.

16. *Кротов В.Ф.* (ред.). Основы теории оптимального управления: учебное пособие. М.: Высшая школа, 1990. 431 с.

17. *Ногин В.Д.* Линейная свертка критериев в многокритериальной оптимизации // Искусственный интеллект и принятие решений. 2014. №. 4. С. 73-82.

18. *Хохлюк В.И.* Параллельные алгоритмы целочисленной оптимизации. М.: Радио и связь, 1987.

MULTI-OBJECTIVE OPTIMIZATION OF VIRTUAL MACHINE PLACEMENT ON PHYSICAL SERVERS IN CLOUD DATA CENTERS

Andrew V. Toutov, *MTUCI, Moscow, Russia,* ***andrew_vidnoe@mail.ru***
Natalia V. Toutova, *MTUCI, Moscow, Russia,* ***e-natasha@mail.ru***
Anatoly S. Vorozhtsov, *MTUCI, Moscow, Russia,* ***as.vorojcov@mail.ru***
Iliya A. Andreev, *MTUCI, Moscow, Russia,* ***1c@mtuci.ru***

Abstract
The problem of virtual machine placement on physical servers in cloud data centers is considered. The resource management system has a two-level architecture consisting of global and local controllers. Local controllers analyze the state of the physical servers on which they are located and determine possible underloading, overloading, and overheating states based on the forecast for the next observation window. If one of the listed states is detected, the local controller notifies the global controller, which selects the destination servers to host the virtual machines via migration. It is proposed to place virtual machines based on the criteria of minimum remaining unused resources and violation of SLA agreements. A mathematical formulation of the optimization problem is given, which is equivalent to the known main assignment problem in terms of structure, necessary conditions, and the nature of variables. Reducing the assignment problem to a closed transport problem allowed us to solve the problem of hosting virtual machines under many criteria in real time and significantly increase its dimension in comparison with heuristic algorithms, which makes it possible to maintain the quality of modern cloud services in the conditions of rapid growth of physical and virtual resources of data centers. The developed mathematical formulation of the problem and the results of computational experiments can be included in the mathematical software of virtual machine live migration.

Keywords: *virtualization, virtual machines, assignment problem, data center, optimal placement, multi-criteria optimization*

References

1. Vorozhtsov A.S., Tutova N.V., Tutov A.V. (2016), Dynamic computing resource allocation in data centers. *T-Comm*. Vol. 10, No.7. P. 47-51. (in Russian)
2. Xu, J., & Fortes, J. A. (2010, December), Multi-objective virtual machine placement in virtualized data center environments. *In 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*. P. 179-188, IEEE.
3. Vorozhtsov A.S., Tutova N.V., Tutov A.V. (2015), Optimal cloud servers placement in data centers. *T-Comm*. Vol 9. No.6. P. 4-8. (in Russian)
4. Camati, R. S., Calsavara, A., & Lima Jr, L. (2014), Solving the virtual machine placement problem as a multiple multidimensional knapsack problem, ICN 2014. 264 p.
5. Ferdaus, M. H., Murshed, M., Calheiros, R. N., & Buyya, R. (2014). Virtual machine consolidation in cloud data centers using ACO metaheuristic, *In European conference on parallel processing*. Springer, Cham. P. 306-317.
6. Mikryukov A.A., Hantimirov R. (2015), Initial resource provisioning in IAAS clouds besed on the analytic hierarchy process. *Statistics and Economics,* 4. P. 184-187. (In Russian)
7. Gulati, A., Holler, A., Ji, M., Shanmuganathan, G., Waldspurger, C., & Zhu, X. (2012), Vmware distributed resource management: Design, implementation, and lessons learned, *VMware Technical Journal,* 1(1). P. 45-64.
8. Xu, J., & Fortes, J. (2011), A multi-objective approach to virtual machine management in datacenters. *In Proceedings of the 8th ACM international conference on Autonomic computing*. P. 225-234.
9. Moges, F. F., & Abebe, S. L. (2019), Energy-aware VM placement algorithms for the OpenStack Neat consolidation framework. *Journal of Cloud Computing,* 8(1), 2.
10. Shaw, R., Howley, E., & Barrett, E. (2019), An energy efficient anti-correlated virtual machine placement algorithm using resource usage predictions. *Simulation Modelling Practice and Theory*, 93. P. 322-342.
11. Alharbi, F., Tian, Y. C., Tang, M., Zhang, W. Z., Peng, C., & Fei, M. (2019), An ant colony system for energy-efficient dynamic virtual machine placement in data centers. *Expert Systems with Applications,* 120. P. 228-238.
12. Beloglazov, A., Abawajy, J., & Buyya, R. (2012), Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future generation computer systems,* 28(5). P. 755-768.
13. Murtazaev, A., & Oh, S. (2011), Sercon: Server consolidation algorithm using live migration of virtual machines for green computing, *IETE Technical Review*, 28(3). P. 212-231.
14. Farzai, S., Shirvani, M. H., & Rabbani, M. (2020), Multi-objective communication-aware optimization for virtual machine placement in cloud datacenters, Sustainable Computing: Informatics and Systems, 100374.
15. Tutov A.V. (2018) Models and methods of resources allocation of infocommunication system in cloud data centers. *H&ES Research*. Vol. 10. No 6. P. 100-00.
16. Krotov, V. F. (Ed.). (1990), Fundamentals of optimal control theory: textbook. Higher school. (In Russian)
17. Nogin, V. D. (2014), Linear convolution of criteria in multi-criteria optimization, *Artificial intelligence and decision making*, (4). P. 73-82. (In Russian)
18. Hohlyuk V. I. (1987), Parallel integer optimization algorithms, Radio and communications. (In Russian)

Information about authors:
Andrew V. *Toutov, senior lecturer, MTUCI, Moscow, Russia*
Natalia V. Toutova, *assistant professor, MTUCI, Moscow, Russia*
Anatoly S. Vorozhtsov, *assistant professor, MTUCI, Moscow, Russia*
Iliya A. Andreev, *assistant professor, MTUCI, Moscow, Russia*