

УДК 004.9

DOI: [10.26102/2310-6018/2021.35.4.036](https://doi.org/10.26102/2310-6018/2021.35.4.036)

Интеллектуализация использования аппаратных ресурсов центра обработки данных, оказывающего услуги облачных вычислений

Я.В.Метелкин✉, К.А.Маковий

*Воронежский государственный технический университет,
Воронеж, Российская Федерация
flow101@mail.ru✉*

Резюме: В настоящее время набирает популярность технология облачных вычислений – технология, в рамках которой компьютерные ресурсы центра обработки данных предоставляются пользователю по сети как онлайн-сервис. Спрос стимулирует рост количества и размеров центров обработки данных, оказывающих данные услуги. Пандемия стала причиной перехода многих сервисов в режим онлайн. Многие организации подчеркнули для себя эффективность удаленной работы, в связи с чем выросла распространенность дистанционного обучения. Таким образом, возникает необходимость оптимизации работы ИТ-инфраструктуры поставщиков облачных услуг в целях повышения ее экономичности и экологичности (соответствие концепции Green computing) при сохранении заранее определенного уровня качества обслуживания. Одной из ключевых проблем оказания услуг облачных технологий является оптимальное распределение виртуальных машин на физических серверах. Данной проблемой занимались многие исследователи, и в этой работе проведен анализ существующих подходов к ее решению. Все они основаны на анализе текущей рабочей нагрузки с последующей корректировкой распределения виртуальных машин. В работе предлагается интеллектуализация использования аппаратных ресурсов центра обработки данных, которая заключается в превентивном управлении аппаратными платформами, размещении виртуальных машин, основанном на прогнозировании рабочей нагрузки в будущем. Для апробации предлагаемого подхода использованы возможности фреймворка CloudSim.

Ключевые слова: облачные вычисления, виртуализация, прогнозирование рабочей нагрузки, cloudsim, центр обработки данных, оптимизация

Для цитирования: Метелкин Я.В., Маковий К.А. Интеллектуализация использования аппаратных ресурсов центра обработки данных, оказывающих услуги облачных вычислений.

Моделирование, оптимизация и информационные технологии. 2021;9(4). Доступно по: <https://moitvvt.ru/ru/journal/pdf?id=1008> DOI: 10.26102/2310-6018/2021.35.4.036

Hardware resource use intellectualization of data centers providing cloud-computing services

Y.V.Metelkin✉, K.A.Makoviy

*Voronezh State Technical University,
Voronezh, Russian Federation
flow101@mail.ru✉*

Abstract: Nowadays, cloud computing technology is gaining popularity. That is a technology in which the computer resources of a data center are provided for a user over a network as an online service. Demand stimulates the growth in the number and size of data centers that deliver these services. The pandemic has led to the transition of many services to online mode. Many organizations have noted the

effectiveness of remote work and, therefore, the prevalence of distance learning is growing. Thus, there is a need to optimize the operation of service providers' IT infrastructure in order to increase its cost-effectiveness and environmental friendliness (compliance with the Green computing concept) while maintaining a predetermined level of service quality. One of the key challenges of providing cloud services is the optimal distribution of virtual machines on physical servers. This problem has been studied by many researchers, and in this article the analysis of the existing approaches to its solution is carried out. All of them are based on the current workload examination and then adjusting the allocation of virtual machines. This paper proposes hardware resource use intellectualization of a datacenter, which involves proactive management of hardware platforms, placement of virtual machines, based on the prediction of the workload in the future. To test the outlined approach, we used the capabilities of the CloudSim framework.

Keywords: cloud computing, virtualization, workload forecasting, cloudsim, datacenter, optimization

For citation: Metelkin Y.V., Makoviy K.A. Hardware resource use intellectualization of data centers providing cloud-computing services. *Modeling, Optimization and Information Technology*. 2021;9(4). Available from: <https://moitvvt.ru/ru/journal/pdf?id=1008> DOI: 10.26102/2310-6018/2021.35.4.036 (In Russ).

Введение

Облачные вычисления – широко распространенная технология, предоставляющая ряд преимуществ: удобство сопровождения, безопасность, экономия на цикле обновления оборудования, экономия на затратах на сопровождение ИТ-инфраструктуры. Основным принципом предоставления облачных услуг является предоставление вычислительных ресурсов центра обработки данных (ЦОД) по аналогии с коммунальными услугами.

В зависимости от предоставляемых услуг можно выделить три типа данной технологии: «Инфраструктура как услуга» (IaaS), «Платформа как услуга» (PaaS) и «Программное обеспечение как услуга» (SaaS).

По данным Gartner, к 2022 году 60 % организаций будут использовать внешнего поставщика облачных услуг [1], а к 2024 году одни только облачные вычисления будут составлять 14,2 % от общих глобальных расходов на ИТ.

Для бесперебойного предоставления облачных услуг используются массивные сетевые инфраструктуры, а ЦОДы являются ядром этих сетевых систем. И ввиду большого спроса на облачные вычисления количество и размер ЦОДов стремительно растет, а их число в мире составляет более восьми миллионов, с ежегодным приростом 12 %. ЦОДы потребляют 2 % произведенной электроэнергии в мире. Таким образом, возникает необходимость оптимизации работы центра обработки данных, оказывающих облачные вычисления. Задача данной работы – предложить подход к размещению виртуальных машин (ВМ) ЦОДа, позволяющий минимизировать энергетические и экономические остатки, отличающийся от существующих возможностью проактивного управления аппаратными платформами.

Актуальность работы

Оптимизация организации виртуальной ИТ-инфраструктуры – большая задача, содержащая в себе множество этапов. Начинается все с технико-экономического обоснования [2], задачей которого является показать то, каким образом можно виртуализировать рабочие станции и обосновать выгоду. Следующий шаг – выбор аппаратных платформ в зависимости от потребностей [3] потенциальных пользователей. Перед принятием решения о полноценном вводе инфраструктуры в промышленную

эксплуатацию разума реализация пилотного проекта, который позволит учесть все сильные и слабые стороны технического решения [4].

Далее возникает проблема оптимального использования ресурсов серверного парка, минимизации затрат на сопровождение, повышение экологичности вычислений. Ключом ее решения является разумное размещение ВМ, отключение простаивающих серверов.

В литературе описываются различные методы и алгоритмы решения проблем размещения виртуальных машин, но большинство из них являются эвристическими и не гарантируют эффективность в применении.

Авторы Соловьев и Удовиченко в статье [5] предлагают комбинированный метод планирования размещения группы ВМ с перераспределением ресурсов, состоящий из двух уровней оптимизации: глобальной оптимизации на уровне всего множества размещаемых ВМ и локальной оптимизации на уровне планирования размещения одной ВМ. На глобальном уровне выполняется решение многокритериальной задачи оптимизации, определяющей порядок размещения ВМ, с применением жадного алгоритма. На локальном используется метод динамического программирования для определения оптимальной последовательности перераспределения ресурсов с точки зрения времени размещения группы ВМ.

В алгоритме [6] авторы, помимо алгоритма размещения (MBFD), экспериментировали с различными способами определения набора виртуальных машин для миграции.

Основная идея состоит в том, чтобы удалить все виртуальные машины из недостаточно загруженных физических серверов, чтобы их можно было отключить, и удалить некоторые виртуальные машины из перегруженных физических серверов в целях снижения нагрузки. Для этого необходимо определить, когда следует считать сервер перегруженным и какие виртуальные машины следует удалить из перегруженного физического сервера. Авторы обнаружили, что наилучшие результаты достигаются, если обнаружение перегрузки выполняется с помощью локальной регрессии, а для уменьшения перегрузки выбираются виртуальные машины с наименьшим расчетным временем миграции, пока физическая машина больше не будет перегружена.

Распределитель, который описывается в статье [7], задает сложный набор правил выбора физического сервера для текущей ВМ: основным критерием является выбор физических серверов с наибольшей энергоэффективностью, где энергоэффективность – это отношение мощности процессора к пиковому энергопотреблению, выбирается сервер с наименьшим потреблением энергии, физическая машина с самой высокой загрузкой и емкостью ЦП.

Авторами статьи [8] были рассмотрены несколько алгоритмов упаковки. Из всех представленных алгоритмов, алгоритмы PercentageUtil (Shi – PU) и AbsoluteCapacity (Shi – AC) показали наиболее приемлемыми. В обоих случаях физические серверы сортируются в соответствии с заданной метрикой (использование и емкость, соответственно), а затем алгоритм итеративно пытается освободить сервер с наименьшими показателями, пытаясь перенести его виртуальные машины к серверу с самыми большими показателями.

Алгоритм [9], названный обратным спекулятивным размещением (Backward Speculative Placement, BSP), немного отличается от схем других алгоритмов, потому что он предварительно проверяет возможность миграции всех виртуальных машин наиболее загруженного сервера ко всем другим серверам, в отличие от выбора первого подходящего.

Все эти подходы не учитывают то, какая нагрузка будет в будущем и в какой момент времени, опираясь лишь на текущие показатели. Алгоритмы размещения ВМ можно улучшить, научившись предсказывать рабочую нагрузку, что позволит проактивно использовать аппаратные ресурсы ЦОДа, осуществлять миграцию ВМ, отключать ненужные серверы. Для любого заказчика облачных услуг можно выделить несколько типовых пользователей в зависимости от потребностей в ресурсах, набора используемых приложений, инфраструктуры. Еще одна особенность – это циклический характер нагрузки на центр обработки данных. Наличие этих закономерностей позволяет добиться хороших результатов при использовании методов прогнозирования.

Теоретическая часть

Одним из способов предсказания нагрузки является использование наивного байесовского классификатора. Выделяется несколько классов нагрузки (типов пользователей). Предположим, что для каждого класса $\omega_i \in W$ априорной вероятностью появления класса ω_i будет являться $P(\omega_j)$, а плотность распределения будет равняться $P(x|\omega_j)$.

Тогда, согласно теореме Байеса, вероятность того, что нагрузка x будет принадлежать классу ω_i , вычисляется следующим образом:

$$P(\omega_j|x) = \frac{P(x|\omega_j)P(\omega_j)}{P(x)}. \quad (1)$$

Цель этой классификации состоит в том, чтобы определить, к какому классу будет принадлежать рабочая нагрузка в следующем временном интервале. Наиболее вероятным классом ω^* будет тот, для которого условная вероятность принадлежности максимальна.

Для решения задачи необходимо рассчитать вероятность для всех классов и выбрать тот, для которого вероятность имеет максимальное значение.

Более гибкий метод – это использование нейросетевого подхода (НС) для прогнозирования количества виртуальных машин определенного типа в любой момент времени.

Процент загрузки сервера на каждом последующем шаге зависит от количества аппаратных ресурсов, используемых на предыдущем, поэтому мы предлагаем использовать сеть NARX – модель временных рядов, в которой значения временных рядов в данный момент зависят от предыдущих значений того же ряда.

В этой модели значение выходного сигнала вычисляется на основе значений этого сигнала, полученных в предыдущие моменты времени, а также значений управляющего входного сигнала в предыдущих выборках времени $y(t)$ и $u(t)$. Управляющее уравнение для модели нейронной сети второго типа:

$$y'(t) = f(y(t-1), \dots, y(t-dy), u(t-1), \dots, u(t-du)), \quad (2)$$

где $y'(t)$ – прогнозируемое значение $y(t)$, dy и du представляют собой временную задержку ввода и вывода соответственно. Модель NARX обеспечивает лучший прогноз, чем другие модели NN, поскольку она использует дополнительную информацию, содержащуюся в предыдущих значениях $u(t)$.

Полученный прогноз используется для принятия решения о необходимости перераспределения виртуальных машин.

Пусть $D = \{D_1, D_2 \dots D_n\}$ – набор типов аппаратных ресурсов, $R = \{R_1, R_2 \dots R_n\}$ – прогнозируемый объем ресурса D_{ik} в начале интервала времени k , $S = \{S_1, S_2 \dots S_m\}$ – набор аппаратных платформ ЦОДа, $U = \{U_1, U_2 \dots U_n\}$ – объем ресурса D_j на сервере S_i . Один из критериев оптимизации – это обеспечение уровня SLA – соглашения об уровне

сервиса. Цель оптимизации – обеспечение максимальной доступности по SLA-договору – гарантированного времени предоставления ресурсов за период. Доступность – отношение разности согласованного времени работоспособности и суммарного времени простоя к согласованному времени работоспособности. Ключевым элементом обеспечения доступности является контроль уровня задержки.

Пусть L – это предельный уровень задержки при предоставлении ресурсов в момент времени k . Данная задержка зависит от узкого места, которое определяется ресурсом с наибольшим использованием в каждый интервал k :

$$M = \min(\sum_{i=1}^m U_{ij} - R_j), \quad (3)$$

где $j = 1..n$.

Нарушение SLA:

$$La = q_j(M_j) - L, \quad (4)$$

где $q(M)$ – средняя задержка при текущем использовании ресурса D_j .

Общее энергопотребление аппаратной платформы определяется линейной функцией использования ресурсов процессора, физической и оперативной памяти:

$$E_k = \sum_{i=1}^m (Eidle_i + \sum_{j=1}^n e_j * R_{ij}), \quad (5)$$

где $Eidle_i$ – энергопотребление сервера S_i в состоянии простоя, e_j – потребление единицы ресурса D_j , R_{ij} – прогнозируемое использование ресурса D_j на сервере S_i .

Таким образом, необходимо решить задачу оптимизации:

$$E_k \rightarrow \min, La \rightarrow \min \quad (6)$$

При ограничении:

$$\sum_{i=1}^m \sum_{j=1}^n U_{ij} \geq R_j. \quad (7)$$

В ходе обучения удалось достичь $MSE = 1.08554 \cdot 10^{-4}$. Данной точности достаточно для выбора числа серверов и их конфигурации для обеспечения бесперебойного выполнения всех выполняемых виртуальных машин. На Рисунке 1 показан результат прогнозирования рабочей нагрузки.

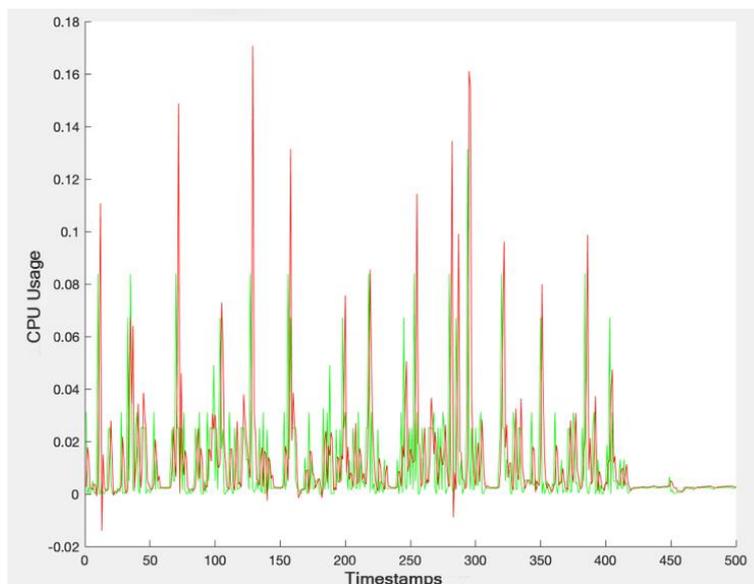


Рисунок 1 – Результаты прогнозирования рабочей нагрузки
Figure 1 – Workload forecasting results

Практическая часть

Для тестирования контроллера распределения был проведен численный эксперимент с использованием фреймворка CloudSim [10], в ходе которого моделируется работа ЦОДа с заданной конфигурацией. Моделируемая рабочая нагрузка является входным параметром для разработанного в рамках данной работы планировщика, который, основываясь на прогнозируемой нагрузке, в будущем осуществляет миграцию ВМ в целях повышения энергоэффективности. Данный планировщик был реализован в рамках класса VmScheduler – класса, который определяет и реализует политику, используемую для совместного использования вычислительных мощностей виртуальными машинами, работающими на указанном хосте.

Проведен численный эксперимент, позволяющий оценить работу планировщика размещения виртуальных машин. Пример использования ресурсов центра обработки данных, имеющего x86 архитектуру, операционную систему Linux, гипервизор Xen, состоящего из четырех хостов, имеющих в сумме 16 ядер, 38400 MIPS, 13ТВ объем постоянной памяти, 140Гб оперативной памяти показан на Рисунках 8-10. В ходе моделирования осуществляется запуск 37 виртуальных машин с различными характеристиками (Таблица 1).

Таблица 1 – Характеристики виртуальных машин

Table 1 – Characteristics of virtual machines

Кол-во, шт.	Размер образа, Мб	Кол-во ядер, шт.	MIPS	RAM, Мб	Гипервизор
10	1000	3	2050	1216	Xen
5	1500	2	1050	2048	Xen
1	1000	1	1000	512	Xen
1	1000	1	1000	1024	Xen
10	1000	1	1000	512	Xen

На каждой виртуальной машине происходит выполнение различных задач, число которых в минуту в среднем составляет 100 штук. На Рисунках 2 и 3 показано использование ресурсов каждого хоста (загрузка процессора, оперативной памяти и сети). Точность прогноза основывается на том, что в рамках недельного цикла, а также в рамках одного дня, нагрузка носит циклический характер, отображающий особенности рабочего процесса в разрезе предметной области.

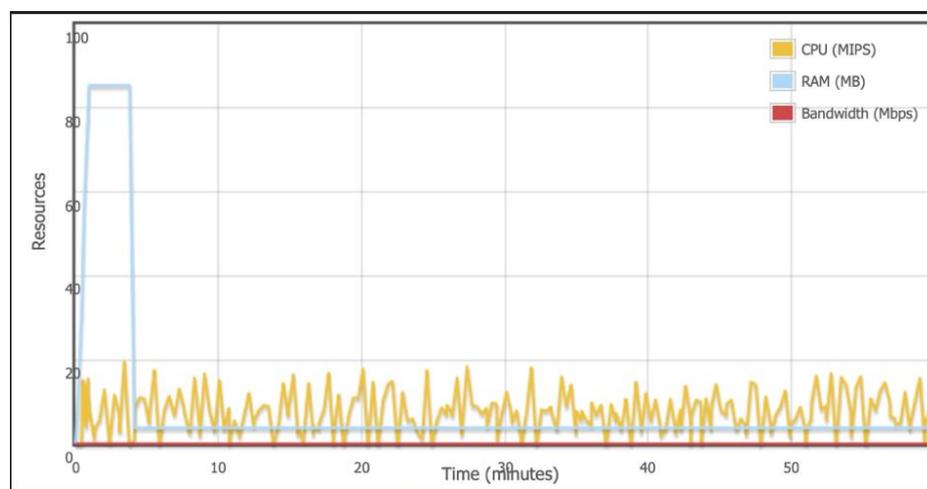


Рисунок 2 – Потребление ресурсов хоста 0

Figure 2 – Host 0 Resource consumption

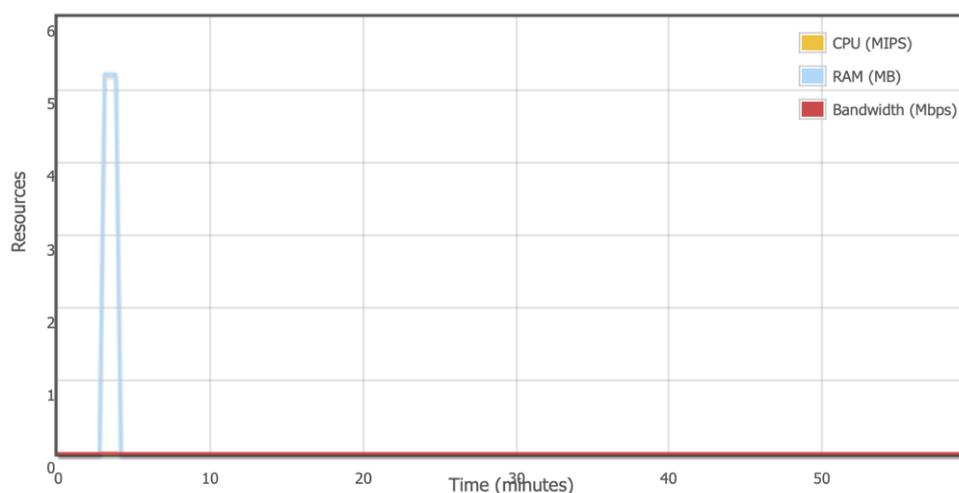


Рисунок 3 – Использование ресурсов хоста 1
Figure 3 – Resource usage of Host 1

Таким образом, предложенная стратегия размещения виртуальных машин позволяет обеспечивать эффективное использование аппаратных ресурсов, проводить гибкую настройку политики, что позволяет достичь целей оптимизации использования облачной инфраструктуры.

Заключение

В работе описаны существующие решения по оптимизации процесса распределения ВМ по ЦОДам. Основной минус данных методов – они основываются на текущем потоке клиентов, результатом чего являются задержки в распределении, появление простаивающих серверов, а простаивающий сервер потребляет до 70 % от своей пиковой мощности. Нами предложен прогностический подход к распределению, позволяющий проактивно управлять аппаратными ресурсами ЦОДа.

СПИСОК ИСТОЧНИКОВ

1. Gartner, Gartner forecasts worldwide public cloud revenue to grow 17% in 2020. Доступно по: <https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020> (дата обращения: 24.04.2021).
2. Makoviy K., Khitskova Y. Estimating the Cost of Implementing Virtual Desktops as a Stage of Project Management in the Field of Cloud Technologies. *International Russian Automation Conference. Springer, Cham.* 2019;641:1034–1043.
3. Маковий К.А., Шипилов Н.В. Анализ потребностей виртуальной машины в ресурсах сервера VDI. *Актуальные проблемы прикладной математики, информатики и механики: сборник трудов международной научно-технической конференции, Воронеж, 12-15 сентября 2016 года.* Воронеж: Научно-исследовательские публикации. 2016:100–103.
4. Маковий К.А., Хищкова Ю.В., Герус С.В. Использование метода гибридных оценок в области информационных технологий. *Научный вестник Воронежского государственного архитектурно-строительного университета. Серия: Информационные технологии в строительных, социальных и экономических системах.* 2016;1(7):120–124.

5. Соловьев В.П., Удовиченко А.О. Метод планирования размещения группы виртуальных машин с перераспределением ресурсов. *Программные продукты и системы*. 2012;1:134-137.
6. Beloglazov A., Buyya R. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience*. 2012;24(13):1397–1420.
7. Lago D.G., Madeira E.R. M., Bittencourt L.F. Power-aware virtual machine scheduling on clouds using active cooling control and DVFS. In Proceedings of the 9th International Workshop on Middleware for Grids. *Clouds and e-Science*. 2011:1–6.
8. Shi L., Furlong J., Wang R. Empirical evaluation of vector bin packing algorithms for energy efficient data centers. In *IEEE Symposium on Computers and Communications*. 2013:9–15.
9. Calcavecchia N., Biran O., Hadad E., Moatti Y. VM placement strategies for cloud scenarios. In *IEEE 5th International Conference on Cloud Computing*. 2012:852–859.
10. Calheiros, R.N., Ranjan, R., Beloglazov, A., De Rose, C.A., & Buyya, R. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience*. 2011;41(1):23–50.

REFERENCES

1. Gartner, Gartner forecasts worldwide public cloud revenue to grow 17% in 2020. Available from: <https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020> (accessed 24.04.2021).
2. Makoviy K., Khitskova Y. Estimating the Cost of Implementing Virtual Desktops as a Stage of Project Management in the Field of Cloud Technologies. *International Russian Automation Conference*. Springer, Cham. 2019;641:1034–1043.
3. Makovij K. A., Shipilov N.V. Analiz potrebnostej virtual'noj mashiny v resursah servera VDI. *Aktual'nye problemy prikladnoj matematiki, informatiki i mekhaniki: sbornik trudov mezhdunarodnoj nauchno-tehnicheskoy konferencii, Voronezh, 12-15 sentyabrya 2016 goda*. Voronezh: Nauchno-issledovatel'skie publikacii. 2016:100–103. (In Russ.)
4. Makovij K. A., Hickova YU.V., Gerus S.V. Ispol'zovanie metoda gibridnyh ocenok v oblasti informacionnyh tekhnologij. *Nauchnyj vestnik Voronezhskogo gosudarstvennogo arhitekturno-stroitel'nogo universiteta*. Seriya: Informacionnye tekhnologii v stroitel'nyh, social'nyh i ekonomicheskikh sistemah. 2016;1(7):120–124. (In Russ.)
5. Solov'ev V. P., Udovichenko A. O. Metod planirovaniya razmeshcheniya gruppy virtual'nyh mashin s pereraspredeleniem resursov. *Programmnye produkty i sistemy*. 2012;1: 134-137. (In Russ.)
6. Beloglazov A., Buyya R. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers. *Concurrency and Computation: Practice and Experience*. 2012;24(13):1397–1420.
7. Lago D.G., Madeira E.R. M., Bittencourt L.F. Power-aware virtual machine scheduling on clouds using active cooling control and DVFS. In Proceedings of the 9th International Workshop on Middleware for Grids. *Clouds and e-Science*. 2011:1–6.
8. Shi L., Furlong J., Wang R. Empirical evaluation of vector bin packing algorithms for energy efficient data centers. In *IEEE Symposium on Computers and Communications*. 2013:9–15.
9. Calcavecchia N., Biran O., Hadad E., Moatti Y. VM placement strategies for cloud scenarios. In *IEEE 5th International Conference on Cloud Computing*. 2012:852–859.

10. Calheiros, R.N., Ranjan, R., Beloglazov, A., De Rose, C.A., & Buyya, R. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience*. 2011;41(1):23–50.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Метелкин Ярослав Викторович, аспирант, Воронежский государственный технический университет, Воронеж, Российская Федерация. *e-mail:* flow101@mail.ru
ORCID: [0000-0003-0317-0948](https://orcid.org/0000-0003-0317-0948)

Metelkin Yaroslav Victorovich, postgraduate student, Voronezh state technical university, Voronezh, Russian Federation.

Маковий Катерина Александровна, старший преподаватель, Воронежский государственный технический университет, Воронеж, Российская Федерация. *e-mail:* makkatya@mail.ru
ORCID: [0000-0003-3921-6047](https://orcid.org/0000-0003-3921-6047)

Makoviy Katerina Aleksandrovna, senior lecturer, Voronezh state technical university, Voronezh, Russian Federation.

Статья поступила в редакцию 16.08.2021; одобрена после рецензирования 21.12.2021; принята к публикации 26.12.2021.

The article was submitted 16.08.2021; approved after reviewing 21.12.2021; accepted for publication 26.12.2021.