ОРГАНИЗАЦИЯ ЭЛАСТИЧНЫХ СИСТЕМ ВИРТУАЛЬНЫХ ОБЛАЧНЫХ СЕРВЕРОВ

Дос Е.В.[1], Камалиденов К.Ш.[2], Мостовщиков Д.Н.[3]

[1]Дос Евгений Владимирович - старший системный архитектор,
Li9, Inc., г. Феникс, Соединенные Штаты Америки;
[2]Камалиденов Куаныш Шарипханович - ведущий системный архитектор,
Digital IQ, г. Нур-Султан, Республика Казахстан;
[3]Мостовщиков Дмитрий Николаевич - старший системный архитектор,
Li9, Inc., г. Феникс, Соединенные Штаты Америки

Аннотация: *рассмотрены методы организации информационных систем облачных сервисов как комплексов виртуальных машин. Предложена базовая модель эластичной мультисерверной системы, которая включает алгоритмы горизонтального и вертикального автомасштабирования. Для проведения численной оценки продуктивности и ресурсоемкости эластичной мультисерверной системы было предложено построить математическую модель системы массового обслуживания, целевыми функциями которой являются показатели времени обработки запросов и требований к вычислительной мощности системы. Для построения эластичной модели мультисерверной системы были определены особенности эластичных систем массового обслуживания и было предложено использовать модель непрерывной цепи Маркова.*
В результате проведенного исследования была построена математическая модель горизонтального автомасштабирования системы массового обслуживания.
Ключевые слова: *информационная система, виртуальная машина, облачные вычисления, эластичность системы, автомасштабирование системы, система массового обслуживания, модель непрерывной цепи Маркова.*

ORGANIZATION OF ELASTIC VIRTUAL CLOUD SERVER SYSTEMS

Dos E.V.[1], Kamalidenov K.Sh.[2], Mostovshchikov D.N.[3]

[1]Dos Evgenii Vladimirovich - Senior Systems Architect,
LI9, INC., PHOENIX, AZ, UNITED STATES OF AMERICA;
[2]Kamalidenov Kuanysh Sharipkhanovich - Senior Systems Architect,
DIGITAL IQ, NUR-SULTAN, REPUBLIC OF KAZAKHSTAN;
[3]Mostovshchikov Dmitrii Nikolayevich – Senior Systems Architect,
LI9, INC., PHOENIX, AZ, UNITED STATES OF AMERICA

Abstract*: the methods of organizing of cloud services information systems as complexes of virtual machines are considered. A basic model of an elastic multi-server system is proposed, which includes algorithms for horizontal and vertical automatic mode scaling. To conduct a numerical evaluation of the productivity and resource consumption of an elastic multi-server system, it was proposed to construct a mathematical model of a queuing system, the target functions of which are indicators of query processing time and requirements for the computing power of the system. To build an elastic model of a multi-server system, the features of elastic queuing systems were determined and it was proposed to use the continuous-time Markov chain model. As a result of the study, a mathematical model of horizontal automatic mode scaling of the queuing system was built.*
Keywords: *information system, virtual machine, cloud computing, system elasticity, system automatic mode scaling, queuing system, continuous-time Markov chain model.*

УДК 004.75

Введение

Современные информационные системы (ИС) характеризуются эффективной организацией автоматического доступа к серверам, рабочим станциям, репозиториям данных, узлам глобальных и локальных сетей, программным приложениям и сервисам и т. д. [1, 21]. Это обеспечивается путем настройки производительной работы систем контроля, мониторинга и обработки запросов, расширения каналов сетевой передачи данных, построения жизненного цикла (ЖЦ) масштабируемой системы на уровне проектирования ИС, разработки методов кластеризации и объединения вычислительных ресурсов, а также внедрения в оптимизационную схему анализа инфраструктуры ИС такого фактора, как эластичность (elasticity). Именно эластичность ИС является параметром, соответствующим главному преимуществу облачных сервисов [1–4] перед другими новейшими концепциями, такими как парадигма распределенных вычислений, кластерных вычислений и распределенных вычислений, что определяет актуальность исследования тематики организации систем виртуальных облачных серверов.

Анализ современных исследований и публикаций в указанной области указывает на необходимость решения задачи оптимальной конфигурации мультисервера для максимизации эффективности функционирования облачных вычислительных сред на базе модели очереди M/M/m [1, 2].

Были рассмотрены модель оптимального распределения мощности и распределения нагрузки для многоядерных серверных процессоров облачной платформы [3, 7], методы конкурентной обработки запросов пользователей [16, 17], методики распределения мощности для нескольких гетерогенных серверов [14], погрузка для нескольких гетерогенных серверов [12] и многоядерного серверного процессора [13],

конфигурации многоядерного серверного процессора [11], а также подходы к управлению производительностью потребления энергии для нескольких гетерогенных серверов [23], включая проактивное планирование (proactive scheduling). В связи с тем, что в рамках исследования в качестве эффективного инструмента предлагается использовать модель непрерывной цепи Маркова (CTMC, Continuous-Time Markov Chain), были рассмотрены аналитические модели CTMC-модели [5, 6, 8] и MAP-модели очереди [22].

Анализ задачи внедрения эластичности системы был сосредоточен на методиках динамического управления мощностью, как функции, зависящей от нагрузки [11], и алгоритмах количественного определения параметра эластичности для облачных платформ [15]. Проведенный анализ указал на отсутствие целостной методологии организации ИС облачных серверов, которые могут быть рассмотрены как комплексы виртуальных машин, что является нерешенной частью общего исследования.

Целью работы, таким образом, стало построение комплексной методологии оптимальной организации систем виртуальных облачных серверов на базе CTMC-модели в соответствии с целевой функцией эластичности ИС.

1. Принципы построения эластичной мультисерверной системы

Базовый подход, используемый при построении модели мультисерверной системы облачного сервиса, заключается в представлении серверного комплекса как гомогенного кластера функциональных узлов, что соответствует множеству виртуальных машин (ВМ) с одинаковыми параметрами вычислительной мощности, оперативной памяти и ширины канала сетевой передачи данных. Показатель эластичности при этом указывает на возможность динамического изменения количества узлов или мощности отдельного узла в соответствии с величиной нагрузки, определяемой через количество запросов.

В рамках указанного подхода каждый из узлов, представляющий собой виртуальную машину, рабочую станцию или ядро многоядерного процессора называется сервером, а общая система может быть определена как мультисерверная. При этом параметр скорости отдельного сервера соответствует количеству задач, которые могут быть обработаны данным сервером в единицу времени.

Таким образом, на уровне базовой модели можно отметить, что эластичная мультисерверная система на автоматическом уровне осуществляет масштабирование в соответствии с количеством запросов, поступающих в облачный сервис. Согласно общепринятой классификации можно выделить две группы методов автомасштабирования [1, 6]:

автомасштабирование по размеру мультисерверной системы (горизонтальное масштабирование);

автомасштабирование по скорости серверов (вертикальное масштабирование).

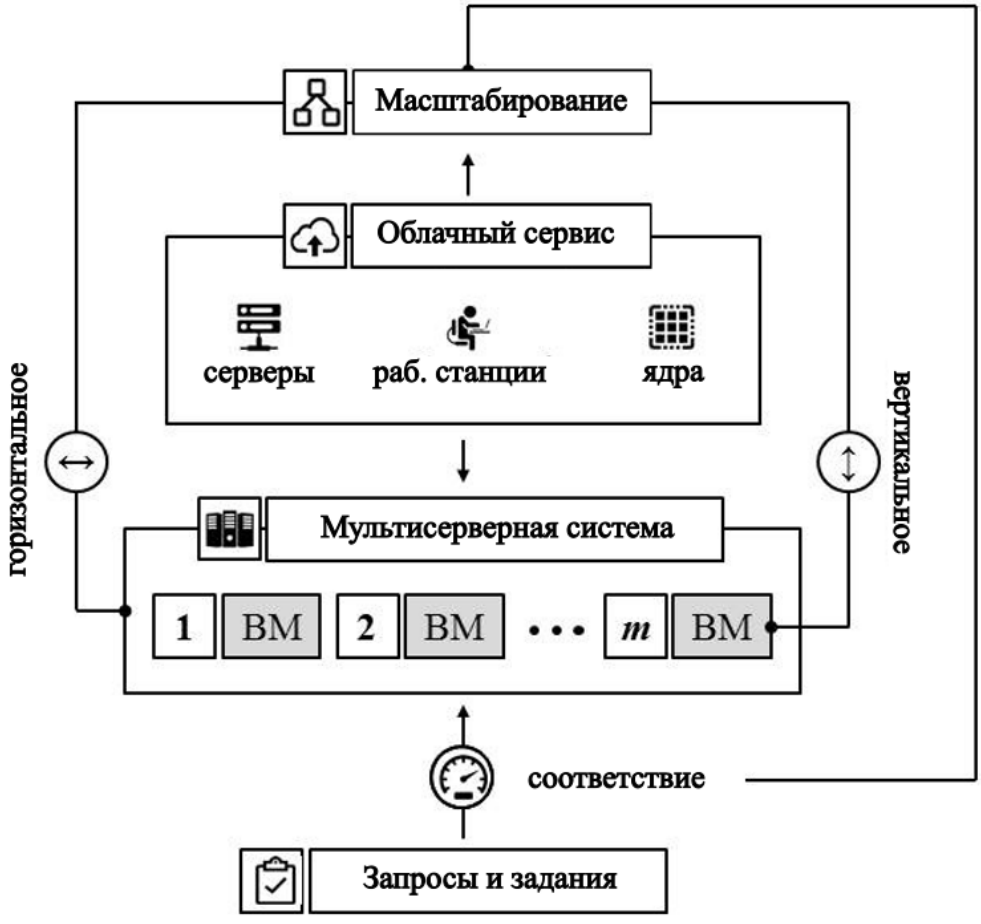

Рис. 1. Базовая модель эластичной мультисерверной системы

Следовательно, при горизонтальном масштабировании в соответствии с изменениями количества запросов изменяется общее количество серверов, а при вертикальном масштабировании – скорость отдельных серверов (рис. 1).

2. Математическое моделирование мультисерверной системы

Для проведения численной оценки производительности и ресурсоемкости эластичной мультисерверной системы нужно выстроить математическую модель системы массового обслуживания (СМО) для облачного сервиса. Целевыми функциями для такой модели будут выступать показатели времени обработки запросов и требования к вычислительной мощности системы, а аргументами целевых функций – переменные, характеризующие вычислительный ресурс серверов и их общее количество в пределах системы, которая моделируется.

СМО в общем случае моделируется через пуассоновский процесс, т.е. запросы рассматриваются как случайные события экспоненциального распределения с постоянным средним значением 1/λ часовых интервалов их регистрации, где λ – интенсивность пуассоновского процесса. В таком случае время обработки запроса может быть определено через переменную t=1/ν, где ν – средняя скорость обработки (среднее количество запросов, которые сервер может выполнить за единицу времени). На основе средней скорости обработки можно рассчитать уровень использования сервера как средний процент времени загрузки:

$$\rho = \frac{\lambda}{m \cdot \nu} \rightarrow \rho = \frac{\lambda \cdot \bar{t}}{m} \qquad (1)$$

Достоверность того, что количество задач составляет N, рассчитывается через PN. Если количество запросов, поступающих в мультисерверную систему, меньше количества серверов, вероятность составляет [19]:

$$P_N = \frac{m^N \rho^N}{n! \cdot \left(\sum_{n=0}^{m-1} \left(\frac{m^n p^n}{n!}\right) + \frac{m^m \rho^m}{m! \cdot (1-\rho)}\right)} для\ N \le m \qquad (2)$$

Аналогично, если количество запросов больше количества серверов, вероятность составляет:

$$P_N = \frac{m^m \rho^N}{m! \cdot \left(\sum_{n=0}^{m-1} \left(\frac{m^n p^n}{n!}\right) + \frac{m^m \rho^m}{m! \cdot (1-\rho)}\right)} для\ N \ge m\,. \qquad (3)$$

Значение Pn для $N \ge m$ может быть использовано при расчете вероятности задержки процесса обработки следующего запроса. Для этого необходимо упростить формулы (1) и (3) путем введения величины ρ0:

$$\rho_0 = \sum_{n=0}^{m-1} \left(\frac{m^n p^n}{n!}\right) + \frac{m^m \rho^m}{m! \cdot (1-\rho)}. \qquad (4)$$

Достоверность задержки процесса обработки нового запроса (queueing probability) к мультисерверной системе является суммой Pn для n∈[m; N]:

$$P_Q = \sum_{n=m}^{N} P_n \rightarrow P_Q = \frac{\rho_0 \cdot m^m \rho^m}{m! \cdot (1-\rho)}\,. \qquad (5)$$

Соответственно, если количество запросов обозначить как K, то среднее значение рассчитывается следующим образом:

$$\overline{K} = \sum_{n=m}^{N} n \cdot P_n \rightarrow \overline{K} = m \cdot \rho + \frac{\rho \cdot P_Q}{1-\rho}\,. \qquad (6)$$

В свою очередь на основе $\overline{K}$ определяется среднее время отклика мультисерверной системы на запрос:

$$\bar{T} = \frac{\overline{K}}{\lambda} \rightarrow \bar{t} + \frac{\bar{t} \cdot P_Q}{m \cdot (1-\rho)} \qquad (7)$$

Разработанная модель СМО предназначена для неэластичной мультисерверной системы, характеризующейся постоянным размером и скоростью серверов, а следовательно, должна быть расширена для построения модели эластичной мультисерверной системы и последующей работы с ней.

3. Горизонтальное автомасштабирование на базе цепей Маркова

Для построения эластичной СМО-модели мультисерверной системы необходимо определить само понятие эластичности [1, 4, 6, 14]. В рамках данной работы мы выделяем следующие ключевые пункты:

Моделирование потока запросов, поступающих в мультисерверную систему в виде пуассоновского процесса, характеризуется скоростью поступления запросов λ как количество запросов на единицу времени. При этом величины временных интервалов между поступлением запросов взаимно независимы и распределены по экспоненциальному закону со средним значением 1/λ.

При полной загруженности серверов мультисерверная система организует очередь запросов, размер которой не ограничен, а обработка новых запросов осуществляется согласно подходу FCFS, то есть в соответствии с местом в очереди первого освободившегося сервера.

Горизонтальное масштабирование осуществляется путем добавления нового сервера. Указанная процедура осуществляется автоматически, время инициализации нового сервера является также экспоненциально распределенной случайной величиной со средним значением 1/μ.

Согласно указанным пунктам можно отметить, что моделирование эластичной мультисерверной системы с горизонтальным автомасштабированием можно осуществить на базе модели СТМС. Как было показано в предыдущем разделе, состояние мультисерверной системы определяется через пару переменных (m, N), где m∈ [1, ∞) и N∈ [1, ∞). Для построения СТМС-модели необходимо ввести дополнительную пару переменных (xm, ym). При этом:

- $m \leq x_m + 1$ для $\forall m \in [1, \infty)$;
- $x_i < x_j$для $\forall i \in [1, m]$ и $\forall j \in [1, m]$, *где* $i < j$;
- $x_i < y_j$для $\forall i \in [1, m]$ и $\forall j \in [1, m]$ *как* для $i < j$, так и для $i > j$.

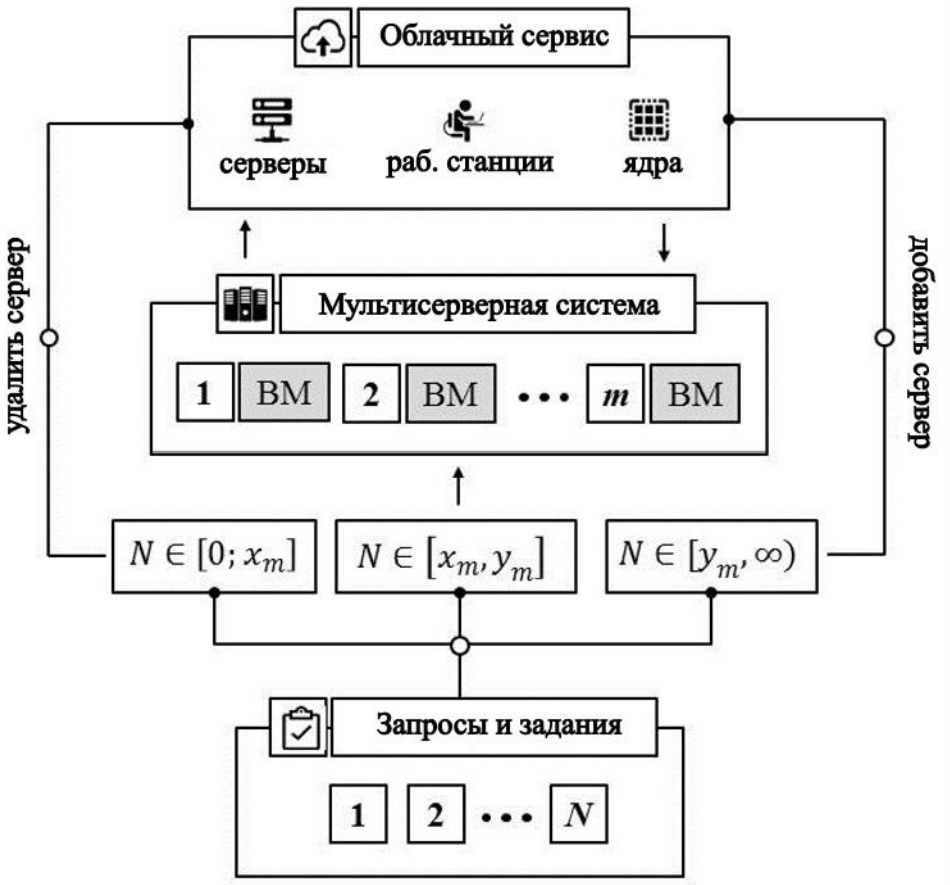

Рис. 2. Математическая модель горизонтального автомасштабирования

Согласно соотнесению N и пары переменных (xm, ym) можно отметить следующие три состояния системы (рис. 2):

$N \in [0; x_m]$ – состояние чрезмерного объема вычислительных ресурсов, при котором можно удалить из системы сервер;

- $N \in [x_m, y_m]$ – нормальное состояние системы;
- $N \in [y_m, \infty)$– состояние недостатка вычислительных ресурсов, при котором необходимо добавить в систему сервер.

Таким образом, предложенную СТМС-модель эластичной мультисерверной системы с горизонтальным автомасштабированием можно представить как ЖЦ функциональных элементов СМО, количество которых составляет m ∈ [1; M], где M – максимальное количество серверов в серверном комплексе. Переход между состояниями системы вычисляется через показатель интенсивности перехода (рис. 3), что характерно для модели цепи Маркова [23].

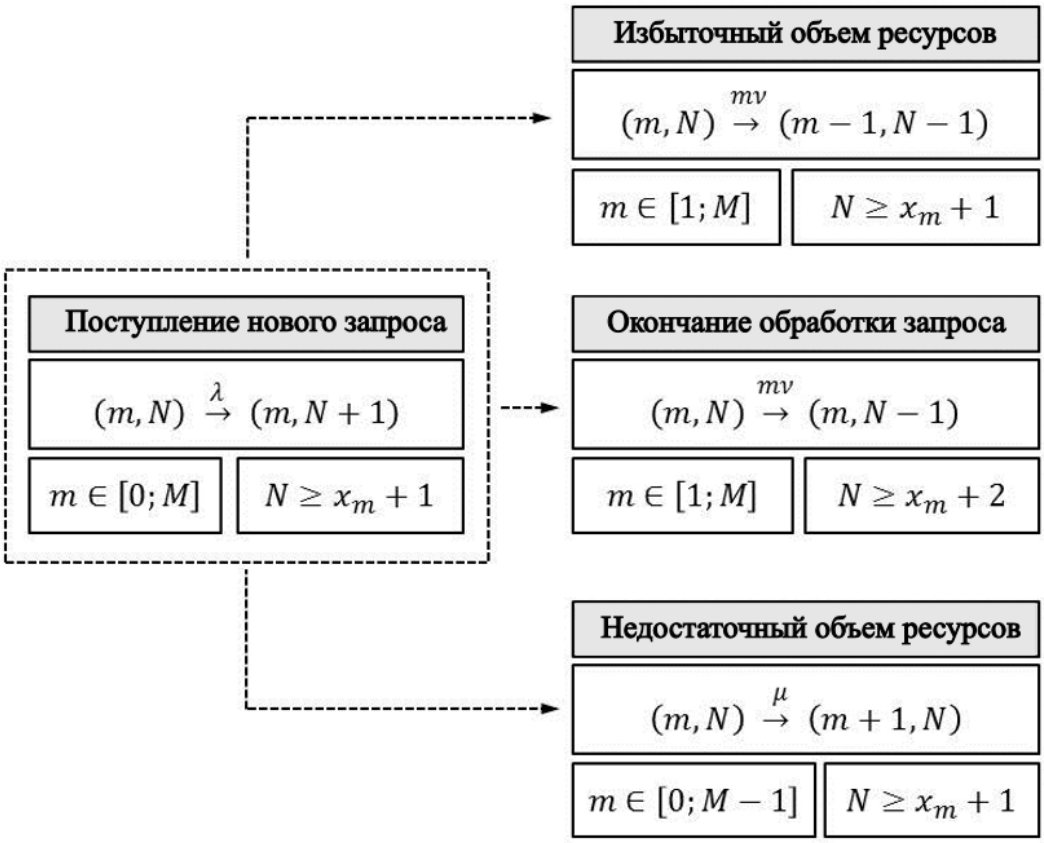

Рис. 3. Схема перехода между состояниями для СМО в рамках СТМС-модели

На основе разработанного математического аппарата можно построить алгоритмы горизонтального автомасштабирования для СМО в рамках СТМС-модели.

Выводы

Таким образом, был проведён анализ современных методов организации информационных систем облачных сервисов, которые могут быть представлены в виде комплексов ВМ. В результате анализа была предложена базовая модель мультисерверной системы, включающая алгоритмы горизонтального и вертикального автомасштабирования. Также для проведения дальнейшей численной оценки продуктивности и ресурсоемкости эластичной мультисерверной системы было предложено построить математическую модель системы массового обслуживания, целевыми функциями которой являются показатели времени обработки запросов и требований к вычислительной мощности системы. Для построения эластичной модели мультисерверной системы были определены особенности эластичных систем массового обслуживания и было предложено использовать модель непрерывной цепи Маркова. В результате проведенного исследования была построена математическая модель горизонтального автомасштабирования системы массового обслуживания, базирующаяся на схеме перехода между состояниями мультисерверной системы.

Список литературы / References

1. *Li K.* (2019). Analytical Modeling and Optimization of an Elastic Cloud Server System From Parallel to Emergent Computing, 31–48. doi: 10.1201/9781315167084-2.
2. *Cao J., Li K., and Stojmenovic I.* “Optimal power allocation and load distribution for multiple heterogeneous multicore server processors across clouds and datacenters,” IEEE Transactions on Computers. Vol. 63. № 1. Pp. 45–58, 2014.
3. *Cao J., Hwang K., Li K., and Zomaya A.* “Optimal multiserver configuration for profit maximization in cloud computing,” IEEE Transactions on Parallel and Distributed Systems. Vol. 24. № 6. Pp. 1087–1096, 2013.
4. *Albonico M., Mottu J.-M., Sunyé G., Alvares F.* (2017). Making Cloud-based Systems Elasticity Testing Reproducible. Proceedings of the 7th International Conference on Cloud Computing and Services Science. doi: 10.5220/0006308905230530.
5. *Ghosh R., Naik V.K. and Trivedi K.S.* “Power-performance trade-offs in IaaS cloud: A scalable analytic approach,” IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops. Pp. 152–157. Hong Kong. China. 27-30 June, 2011.
6. *Martemyanov Y.P., Matveenko V.D.* (2014). On the dependence of the growth rate on the elasticity of substitution in a network. International Journal of Process Management and Benchmarking, 4(4). 475. doi: 10.1504/ijpmb.2014.065524.

7. *Huang J., Li R., Li K., An J. and Ntalasha D.* "Energy-efficient resource utilization for heterogeneous embedded computing systems," IEEE Transactionson Computers. Pp. 1–1, 2017.
8. *Khazaei H., Mišić J., Mišić V.B. and Rashwand S.* "Analysis of a pool management scheme for cloud computing centers," IEEE Transactions on Parallel and Distributed Systems. Vol. 24. № 5. Pp. 849–861, 2013.
9. *Kleinrock Leonard* (2016) Queueing Systems. John Wiley & Sons. New York, 2016.
10. *Li K.* "Improving multicore server performance and reducing energy consumption by workload dependent dynamic power management," IEEE Transactions on Cloud Computing. Vol. 4. № 2. Pp. 122–137, 2016.
11. *Li K.* "Optimal configuration of a multicore server processor for managing the power and performance trade-off," Journal of Supercomputing. Vol. 61. № 1. Pp. 189–214, 2012.
12. *Nourbakhsh V. & Turner J.* (2017). Routing Heterogeneous Jobs to Heterogeneous Servers: A Global Optimization-Based Approach. SSRN Electronic Journal. doi: 10.2139/ssrn.2967811.
13. *Li K., Liu C., Li K.,and Zomaya A.* "A framework of price bidding configurations for resource usage in cloud computing," IEEE Transactions on Parallel and Distributed Systems. Vol. 27. № 8. Pp. 2168–2181, 2016.
14. *Li K.* "Quantitative modeling and analytical calculation of elasticity in cloud computing," IEEE Transactions on Cloud Computing, Vol. 4. Pp. 1–14, 2017.
15. *Liu C., Li K., C.-Z. Xu, and Li K.* "Strategy configurations of multiple users competition for cloud service reservation," IEEE Transactions on Parallel and Distributed Systems. Vol. 27. № 2. Pp. 508–520, 2016.
16. *Li K.* "Optimal power allocation among multiple heterogeneous servers in a data center," Sustainable Computing: Informatics and Systems. Vol. 2. № 1. Pp. 13–22, 2012.
17. *Li K.* "Optimal partitioning of a multicore server processor," Journal of Super computing, vol. 71, № 10, pp. 3744–3769, 2015.
18. *Mei J., Li K. and Li K.* "A fund constrained investment scheme for profit maximization in cloud computing," IEEE Transactions on Services Computing. P. 1–1, 2016.
19. *Mei J., Li K., Ouyang A. and Li K.* "A profit maximization scheme with guaranteed quality of service in cloud computing," IEEE Transactions on Computers. Vol. 64. № 11. Pp. 3064–3078, 2015.
20. *Mei J., Li K. and Li K.* "Customer-satisfaction-aware optimal multi server configuration for profit maximization in cloud computing," IEEE Transactions on Sustainable Computing, Vol. 2. № 1. Pp. 17–29, 2017.
21. *Metheny M.* (2017). Applying the NIST risk management framework. Federal Cloud Computing, 117–183. doi: 10.1016/b978-0-12-809710-6.00005-6.
22. *Melland P., Grance T.* "The NIST definition of cloud computing," Special Publication 800-145, National Institute of Standards and Technology, U.S. Department of Commerce, September, 2011.
23. *Qasmi F., Shehab M., Alves H. & Latva-Aho M.* (2019). Fixed Rate Statistical QoS Provisioning for Markovian Sources in Machine Type Communication, 2019 16th International Symposium on Wireless Communication Systems (ISWCS). doi: 10.1109/iswcs.2019.8877299.
24. *Tian Y., Lin C. and Li K.* "Managing performance and power consumption trade off for multiple heterogeneous servers in cloud computing" Cluster Computing. Vol. 17. № 3, Pp. 943–955, 2014.