Capacity Planning of Cloud Computing Workloads

A Systematic Review

Carlos Diego Cavalcanti Pereira
CESAR – Recife Center for Advanced Studies and Systems
Recife, Brazil
Email: cdcp@cesar.org.br

Felipe Silva Ferraz
CESAR – Recife Center for Advanced Studies and Systems
Recife, Brazil
Email: fsf@cesar.org.br

***Abstract*—Cloud Computing is a prominent field of research with several areas of knowledge to be explored. The current state of the art of cloud computing regarding capacity planning is a more specific field to address in research and further studies. This work has the objective of identifying, evaluating and interpreting published research that examines sizing and capacity planning for cloud computing workloads. To achieve that, a systematic literature review was conducted. This review resulted in the finding of 504 works, of which 52 were identified as primary studies. The studies were then classified according to research focus and aspect of cloud capacity planning. The work investigates what is known about capacity planning models for cloud computing workloads. The results show statistical data about cloud capacity planning, gaps in current research and models for sizing cloud computing workloads with no historical use and workloads based on functional characteristics or architecture.**

***Keywords - cloud capacity planning; capacity planning; cloud computing*.**

I. INTRODUCTION

Cloud Computing is a way of referring to the use of shared computing resources [1]. Cloud Computing groups gather a large number of servers and other computing resources and generally offer combined capacity based on payment on demand and by cycle [2]. Conceptually, Cloud Computing deals equally with partial or complete abstraction of computational capacity, delivering infrastructure components in the form of service to the end customer [3].

Cloud Computing brought us a new paradigm after the evolution of the use of mainframes for x86 servers [4]. In this model, users no longer have control over the physical technology infrastructure [5]. Cloud computing describes a new approach for how computing services and components are available to users.

One of the key aspects to define and implement a cloud computing workload is to understand the appropriate amount of resources needed to meet demand. Mainly, this activity is conducted by applying empirical approaches [6]. On the other hand, "empirical methods" generally imply that the workloads use some sort of historical data to address the sizing – which is not always possible, especially in innovative systems. Another aspect is that to define the amount of resources needed by a specific workload, it is important to understand its architecture, since, even though historical data may be available, it is not effective to assume that this workload has appropriate enhancements in terms of the amount of resources needed to meet the demand.

To understand how those gaps are usually managed, a Systematic Literature Review (SLR) was conducted to map out how to address those issues and processes with regards to capacity planning of cloud computing workloads.

This work is structured as follows: Section I presents the introduction of Capacity Planning of Cloud Computing Workloads; in Section II, related works are presented; in Section III, a brief introduction to cloud capacity planning is addressed; in Section IV, the applied protocol of this systematic review is presented; in Section V, all results are presented; in Section VI, all findings are addressed and discussed; finally, in Section VII, conclusions are presented.

II. RELATED WORK

Considering that Cloud Computing is a relatively new research field, especially in the case of Capacity Planning, there was no related work found regarding Systematic Literature Reviews on this subject. Even so, the different aspects addressed in this research can be found individually in the primary studies found as a result of this Systematic Literature Review (SLR).

In regards of capacity planning models for cloud computing workloads, most of current approaches somehow apply historical use data as key source of information to establish workload resource needs [6]. Although this assertion can be confirmed as presented on the results of this research, there was no research found in the systematic review addressing cloud capacity planning. When evaluating the results of this study, when relating systematic reviews and cloud computing, the only topic addressed in the research found was Cloud Migration [1]. So, is possible to assume that capacity planning of cloud computing workloads is a subject that has unanswered questions that are relevant to be studied.

III. CLOUD CAPACITY PLANNING

Restrictions regarding software development projects, especially considering shortened schedule horizons and contracted time-to-market deadlines, manifest in traditional approaches to capacity planning, where often a gap is seen and is a major risk compromising their production plans [6]. A formal Capacity Planning approach facilitates forecasting

 ISBN: 978-1-61208-827-3

of sizing requirements based on the opportunistic use of whatever performance data and tools are available [1][6][54]-[56]. One of the key aspects when analyzing the relevance of capacity planning in cloud computing projects is the amount of resources needed to meet demand. Depending on the stage at which a project based on cloud computing is, it may be economically unfeasible. This is because architectural decisions can directly impact the need for resources and consequently make the project unviable [4]. Thus in scenarios where there are resource limitations, it is essential to establish a formal capacity planning model [7].

IV. APPLIED PROTOCOL

For the development of this study, general approaches for performing systematic reviews in software engineering [8] and also for its analysis [9] were applied. Our review process has six steps: (1) establish research protocol, (2) inclusion and exclusion criteria definition, (3) perform search (4) content assessment, (5) data extraction, and (6) synthesis.

The objective of this review is to identify current approaches in scientific literature on sizing and capacity planning for cloud computing workloads. The following questions help identify primary studies:

- What are the capacity planning models for cloud computing workloads available in scientific literature?
- Do capacity planning models consider workloads with no historical use?
- Are there capacity planning models for cloud computing workloads based on functional characteristics or its architecture?

A. Inclusion and Exclusion Criteria

For this systematic review, we considered studies that focus in analyzing cloud capacity planning models. The studies could refer to cloud capacity planning specifically or have a broader scope, taking in consideration both cloud computing and capacity planning individually. Considering that this field of research is recent but also in constant development, this review examined studies published from the year 2017.

We also excluded:

- Studies not published in the English language;
- Studies that were unavailable online.

B. Search Strategies

The databases considered in the study are in the list below:

- ACM Digital Library;
- IEEE Xplore;
- ScienceDirect – Elsevier.

To ensure that relevant studies would not be excluded when querying different scientific databases, the search strings were tested on each one of databases to guarantee that it would work for all of them. As a result, a general search string was defined:

1. “cloud capacity planning” OR;
2. “capacity planning” AND “cloud computing” OR;
3. “capacity planning” AND “cloud”.

As mentioned before, the due process of database search and search strings were tested individually on each database until a final statement was defined. The searches were performed between March 2020 and April 2020. The results of each search were summarized and later examined in order to identify duplicity among them. Table 1 presents the number of studies found on each database.

TABLE I. NUMBER OF STUDIES FOUND IN EACH DATABASE

Database	Number of Studies
ACM Digital Library	139
IEEE Xplorer	155
Science Direct	219
Amount of Studies	**513**

C. Studies Selection Process

The papers that were collected in the search process were gathered and added to the Mendeley [61] tool. It was found that there were 9 duplicated works among all databases, resulting in a total of 504 non-duplicated papers. Then, they all had their titles analyzed to determine their relevance and adherence to this study. At this stage, the works that did not have a relationship to capacity planning of cloud computing workloads were eliminated. Papers where the titles were unclear about their relation with the subject of this study were put aside to be analyzed in the next step. At the end of this stage, 365 works were excluded and remaining were 137 items for further analysis of abstracts.

At this stage, all works found previously had their abstracts analyzed. Many were also eliminated due to not conforming to the scope of capacity planning of cloud computing workloads. Papers where it was difficult to determine if they conform to the scope of this study due to the aforementioned reasons were included to be filtered out at a further step. As result of this phase of analysis, 75 papers were excluded, thus remaining were 62 to be analyzed more closely. Table 2 presents the number of studies filtered in each step of selection process.

TABLE II. NUMBER OF STUDIES IN SELECTION PHASE

Phase of Selection Process	Number of Studies
1. Databases Search	504
2. Title Analysis	137
3. Abstract Analysis	62

D. Quality Assessment

After analyzing the search results that did not conform to the scope of this review, we moved on to the quality assessment stage. In this stage, all 62 studies were analyzed, and not only titles or abstracts. In the quality assessment, relevance criteria were established to analyze several aspects regarding each paper selected on prior stages.

To assess the quality of publications, eight questions were defined, based on [9], to support in quality assessment

 ISBN: 978-1-61208-827-3

process. Questions supported the analysis, ensuring that relevance and credibility of all papers were being considered. Of the eight questions raised, the first and last one were used to establish whether the paper was relevant for this review. In this case, both questions were used as final exclusion criteria. The other six questions were useful to determine the quality of papers regarding research methods and other related aspects. In this case, those grades supported a formal quality analysis of publications. The questions were:

1. Does the study examine capacity planning models for cloud computing workloads?
2. Is the study based on formal research methods - not just empirical applications?
3. Are the objectives of the study clearly defined?
4. Is the study context adequately described?
5. Were the methods for data collection used and described correctly?
6. Was the research project adequate to achieve the research objectives?
7. Have the research results been properly validated?
8. Does the study directly contribute to this research?

Of the 62 select studies in prior stages, 52 passed to the stage of synthesis and were thus considered primary studies. In the results section, quality assessment process will be described in detail along with the assessment of the 52 remaining studies.

V. RESULTS

As presented previously, 52 studies were identified [9] – [60] as primary studies. In general, all of them address aspects of this systematic review, whether in terms of scope or research questions.

A. Quantitative Analysis

The research process conducted resulted in 52 primary studies. They were written by 185 authors affiliated to institutions from 19 different countries and were published between 2017 and 2020. A total of 82 different keywords were identified in all papers.

Regarding country of origin, most of the publications were from United States and India (both with eight publications, each, comprising 15% of all primary studies), followed by Brazil (six publications, comprising 12% of all primary studies). United Kingdom had four publications, Australia, China, Finland, Iran and Italy, had three publications, followed by Spain with two publications. Germany, Chile, France, Macedonia, Malaysia, Qatar, Sweden, Taiwan, and Ukraine each had one publication. Considering the various different origins, it can be concluded that capacity planning of cloud computing workloads is a globally widespread topic.

The most common keywords used in selected works, with their respective frequency were: cloud computing (13), capacity planning (8), performance model (5), resource management (5), prediction (4), application (3), performance (3), simulation (3), workload (3), auto-scaling (2), big data (2), quality of service (2), resource provisioning (2), web application (2), workload characterization (2). The first two keywords - cloud computing and capacity planning - reflect exactly the subject of this research.

B. Quality Analysis

As presented before, all primary studies were assessed considering eight quality aspects to ensure their credibility and relevance to this review. The purpose of this analysis was to establish an objective evaluation that all papers selected could actually contribute to the conclusions of this review. To do that, each quality criteria was classified as positive (1) or negative (0).

Table 3 presents the results of this quality assessment of each one of all 52 selected papers. Columns "Q1" to "Q8" represent all of the criteria defined by questions to evaluate the following aspects of publications: Focus, Research, Objectives, Context, Data Collection, Research Project, Validation and Added Value. As mentioned before, all of the selected papers were marked "1" in both "Focus" and "Added Value" criteria. All studies with negatives answers (0) in one of those two criteria were removed during the selection stage.

TABLE III. QUALITY ANALYSIS OF PRIMARY STUDIES

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total
[9]	1	1	1	1	1	0	1	1	88%
[10]	1	1	1	0	1	1	1	1	88%
[11]	1	1	0	1	1	1	1	1	88%
[12]	1	1	1	1	1	1	1	1	100%
[13]	1	1	1	0	1	1	1	1	88%
[14]	1	1	1	0	1	1	1	1	88%
[15]	1	1	0	1	0	1	1	1	75%
[16]	1	1	1	1	0	0	1	1	75%
[17]	1	1	1	1	0	1	1	1	88%
[18]	1	1	0	1	1	1	1	1	88%
[19]	1	1	0	0	1	1	1	1	75%
[20]	1	1	0	1	0	1	0	1	63%
[21]	1	1	0	1	0	1	0	1	63%
[22]	1	1	1	0	0	1	0	1	63%
[23]	1	1	1	0	0	1	1	1	75%
[24]	1	1	0	0	1	1	1	1	75%
[25]	1	1	1	1	0	1	1	1	88%
[26]	1	1	0	1	1	0	1	1	75%
[27]	1	1	0	1	0	1	0	1	63%
[28]	1	1	1	1	1	1	0	1	88%
[29]	1	1	1	1	1	1	1	1	100%
[30]	1	1	0	1	1	0	1	1	75%
[31]	1	1	1	1	1	1	1	1	100%
[32]	1	1	1	1	0	1	1	1	88%
[33]	1	1	1	1	0	1	1	1	88%

 ISBN: 978-1-61208-827-3

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total
[34]	1	1	0	1	0	1	0	1	63%
[35]	1	1	0	1	0	1	0	1	63%
[36]	1	1	0	1	1	1	0	1	75%
[37]	1	1	0	1	1	1	0	1	75%
[38]	1	1	1	0	1	1	1	1	88%
[39]	1	1	1	0	1	1	1	1	88%
[40]	1	1	1	0	0	0	1	1	63%
[41]	1	1	1	1	0	0	1	1	75%
[42]	1	1	1	1	1	0	1	1	88%
[43]	1	1	1	1	1	0	1	1	88%
[44]	1	1	1	1	0	1	1	1	88%
[45]	1	1	1	1	1	1	0	1	88%
[46]	1	1	1	1	1	0	1	1	88%
[47]	1	1	1	1	1	0	1	1	88%
[48]	1	1	1	1	1	0	1	1	88%
[49]	1	1	1	1	0	0	1	1	75%
[50]	1	1	1	1	1	1	1	1	100%
[51]	1	1	1	1	1	1	1	1	100,00%
[52]	1	1	1	1	1	1	1	1	100%
[53]	1	1	1	1	1	1	1	1	100%
[54]	1	1	1	1	1	0	1	1	88%
[55]	1	1	1	0	1	1	0	1	75%
[56]	1	1	1	0	1	1	0	1	75%
[57]	1	1	1	1	1	1	1	1	100%
[58]	1	1	1	1	1	1	1	1	100%
[59]	1	1	1	1	1	1	1	1	100%
[60]	1	1	1	1	1	1	1	1	100%

All of the papers that were analyzed in this review provided information on the research method, adding an important value regarding its relationship with the scientific method. Considering that all studies applied some sort of formal research method, all criteria scored above 70%, except one where data collection scored 67%. Even so, this lack of clarity as to the methods of data collection in some of the works does not generally compromise the quality of the selected papers.

VI. DISCUSSION

After performing the search, data extraction and synthesis of primary studies, the authors were able to identify some patterns regarding capacity planning of cloud computing workloads. At first, it was possible to conclude that cloud computing - and the process of capacity planning for workloads running on cloud - is a very recent field of research and also subject to a lot of entropy, given the characteristic of being fast evolving within computer science.

It is also possible to conclude that there is a lack of standardization of capacity planning methods for cloud computing workloads and methods that are not intensive on historical use data. This was identified after the classification of primary studies in a parallel to research questions - that covers both aspects mentioned previously. The majority of studies applied historical use data to predict future resource demands for a specific type of workload - such as IoT (Internet of Things) solutions, database and so forth - using machine learning and artificial intelligence techniques.

A. Cloud Capacity Planning models

Capacity planning is a process that is not applied only in the field of computer science. Most engineering sciences - or any field that works with limited resources - need to address how to manage and properly apply resources to meet changing and constant evolving demands.

As such, in cloud computing environments, in which resources are offered as services, they are considered as practically infinite - as long as the customer pays for it. Capacity planning models are being applied to manage how to use resources efficiently and in a well architected way.

This research has found that although scientific literature covers formal methods to perform capacity planning for cloud computing workloads, there is no standardization regarding inputs and outputs, processes and generalization, to cover broader scenarios and types of workloads.

B. Cloud Capacity Planning models for cloud computing workloads with no historical use

An important finding of this review was that most of capacity planning methods for cloud computing workloads consider historical data use for understanding demands needs and for planning. This empirical approach shows some efficiency – especially when using historical data as an input for prediction models – but it often fails to deliver a higher percentage of assertiveness on new workloads. Another gap on this type of approach is that when performing capacity planning for an unprecedented type of workload – such as innovative or disruptive software – whereby there is no historical data for that workload; this leads current methods to apply a benchmark as input for those prediction models, decreasing percentage of assertiveness on capacity planning metrics for new or unprecedented workloads.

C. Cloud Capacity Planning models based on type of workloads and architectural characteristic

Scientific literature analyzed in this review showed that there are methods to perform capacity planning for specific types of workloads – such as IoT, database, fog computing and so forth. However, those models vary widely in their method, calculations, and, especially, assertiveness.

In this sense, this systematic review has not found generalist models which could cover capacity planning broadly and which could also consider specific characteristics of different types of workloads. The authors believe that standardization and generalization in the method would enhance scientific evolution for capacity planning of cloud computing workloads.

 ISBN: 978-1-61208-827-3

D. Towards Cloud Capacity Planning

As presented previously, one of the main challenges of working with cloud computing environments is how to properly plan and calculate the amount of resources needed for a specific set of workload. Besides the use of historical data as major input to predict resource demands and the absence of generalist models for capacity planning, our study found another set of challenges:

- Standardization: The lack of standards in gathering and provisioning capacity planning models makes reuse difficult;
- Assertiveness: Although current models deliver some capacity planning metrics, those calculations often fail to deliver a high percentage of assertiveness to define resource needs;
- Generalization: Most of current models address specific types of workloads, and do not cover a more generalist workload based on its architecture, for instance.

VII. CONCLUSION

This systematic review focuses on mapping and identifying studies that aim to establish a formal process for capacity planning of cloud computing workloads. In the search phase, 504 papers were found, of which 52 were classified as primary studies, following applied selection and quality criteria.

All papers were classified considering their focus on answering the research questions. After this stage, a quality analysis was performed to access how the papers addressed eight different quality criteria, as this method was applied to ensure that each study covered formal scientific methods and covered relevant aspects of this systematic review.

In regards to the aspects of capacity planning, the majority of studies covered some type of formal method to perform capacity planning of cloud computing workloads. Most of them focused on historical data use to somehow predict future resource demands. To do that, machine learning and artificial intelligence techniques were generally applied. Another important aspect in parallel to research questions is that no general method or framework was found to cover different type of workloads - although there are methods to perform cloud capacity planning for specific workloads, as mentioned before, each method however establishes a different approach and is focused in analyzing a specific type of workload.

In order to expand the results found and to improve the conclusions of this systematic review, some considerations about the limitations of this study need to be highlighted:

- Perhaps considering a wider period of publications - more than 3 years of publishing - even the great entropy of the subject;
- Apply search strings that include more keywords with terms related to the object of this research, such as “Resource Management”;
- Look for capacity planning challenges in other science and engineering references, given that resource-limited scenarios is a characteristic not only present in computer science.

For future work and further research, it would be important to analyze specifically capacity planning methods that do not apply historical data use - considering that not all software projects have a precedent of use, such as for innovative and disruptive software - and also to cover different types of workloads - since current methods aim to analyze specific types of cloud computing workloads.

REFERENCES

[1] P. Jamshidi, A. Ahmad, and C. Pahl, “Cloud Migration Research: A Systematic Review,” IEEE Trans. Cloud Comput., vol. 1, no. 2, pp. 142–157, 2013, [Online]. Available: https://ulir.ul.ie/bitstream/handle/10344/3656/Jamshid_cloud.pdf?sequence=2.

[2] S. Bhardwaj, L. Jain, and S. Jain, “Cloud Computing : a Study of Infrastructure As a Service (Iaas),” Int. J. Eng., vol. 2, no. 1, pp. 60–63, 2010, [Online]. Available: http://ijeit.org/index_files/vol2no1/CLOUD COMPUTING A STUDY OF.pdf.

[3] L. Wang et al., “Cloud computing: A perspective study,” New Gener. Comput., vol. 28, no. 2, pp. 137–146, 2010, doi: 10.1007/s00354-008-0081-5.

[4] W. Hasselbring and S. Frey, “Model-Based Migration of Legacy Software Systems to Scalable and Resource-Efficient Cloud-Based Applications: The CloudMIG Approach,” First Int. Conf. Cloud Comput. GRIDs, Virtualization Model., no. c, pp. 155–158, 2010.

[5] S. Sheshadhri and R, Nithiya, “Mapping multi-tier architecture into cloud environment using slicing and virtualization,” 40th IRF Int. Conf., pp. 6–10, 2016.

[6] N. Gunther, Guerrilla capacity planning: A tactical approach to planning for highly scalable applications and services. Springer, 2007.

[7] Barbara A. and Kitchenham. Systematic review in software engineering: where we are and where we should be going. In Proceedings of the 2nd international workshop on Evidential assessment of software technologies. Association for Computing Machinery, New York, NY, USA, September, 2012, 1–2. DOI:https://doi.org/10.1145/2372233.2372235

[8] T. Dybå and T. Dingsøyr, “Empirical studies of agile software development: A systematic review,” Inf. Softw. Technol., vol. 50, no. 9, pp. 833–859, 2008, doi: https://doi.org/10.1016/j.infsof.2008.01.006.

[9] W. Iqbal, A. Erradi, and A. Mahmood, “Dynamic workload patterns prediction for proactive auto-scaling of web applications,” J. Netw. Comput. Appl., vol. 124, pp. 94–107, 2018, doi: https://doi.org/10.1016/j.jnca.2018.09.023.

[10] M. Amiri and L. Mohammad-Khanli, “Survey on prediction models of applications for resources provisioning in cloud,” J. Netw. Comput. Appl., vol. 82, pp. 93–113,

2017, doi: https://doi.org/10.1016/j.jnca.2017.01.016.
[11] M. Amiri, L. Mohammad-Khanli, and R. Mirandola, “A sequential pattern mining model for application workload prediction in cloud environment,” J. Netw. Comput. Appl., vol. 105, pp. 21–62, 2018, doi: https://doi.org/10.1016/j.jnca.2017.12.015.
[12] V. de N. Personé and A. Di Lonardo, “Approximating finite resources: An approach based on MVA,” Perform. Eval., vol. 131, pp. 1–21, 2019, doi: https://doi.org/10.1016/j.peva.2018.11.005.
[13] J. O. de Carvalho, F. Trinta, D. Vieira, and O. A. C. Cortes, “Evolutionary solutions for resources management in multiple clouds: State-of-the-art and future directions,” Futur. Gener. Comput. Syst., vol. 88, pp. 284–296, 2018, doi: https://doi.org/10.1016/j.future.2018.05.087.
[14] R. Tolosana-Calasanz, J. Á. Bañares, and J.-M. Colom, “Model-driven development of data intensive applications over cloud resources,” Futur. Gener. Comput. Syst., vol. 87, pp. 888–909, 2018, doi: https://doi.org/10.1016/j.future.2017.12.046.
[15] K.-J. Wang and P. H. Nguyen, “Capacity planning with technology replacement by stochastic dynamic programming,” Eur. J. Oper. Res., vol. 260, no. 2, pp. 739–750, 2017, doi: https://doi.org/10.1016/j.ejor.2016.12.046.
[16] M. Zakarya and L. Gillam, “Modelling resource heterogeneities in cloud simulations and quantifying their accuracy,” Simul. Model. Pract. Theory, vol. 94, pp. 43–65, 2019, doi: https://doi.org/10.1016/j.simpat.2019.02.003.
[17] M. Amiri, L. Mohammad-Khanli, and R. Mirandola, “An online learning model based on episode mining for workload prediction in cloud,” Futur. Gener. Comput. Syst., vol. 87, pp. 83–101, 2018, doi: https://doi.org/10.1016/j.future.2018.04.044.
[18] V. Medel, R. Tolosana-Calasanz, J. Á. Bañares, U. Arronategui, and O. F. Rana, “Characterising resource management performance in Kubernetes,” Comput. Electr. Eng., vol. 68, pp. 286–297, 2018, doi: https://doi.org/10.1016/j.compeleceng.2018.03.041.
[19] M. S. Aslanpour, M. Ghobaei-Arani, and A. Nadjaran Toosi, “Auto-scaling web applications in clouds: A cost-aware approach,” J. Netw. Comput. Appl., vol. 95, pp. 26–41, 2017, doi: https://doi.org/10.1016/j.jnca.2017.07.012.
[20] B. Treynor, M. Dahlin, V. Rau, and B. Beyer, “The calculus of service availability,” Commun. ACM, vol. 60, no. 9, pp. 42–47, Aug. 2017, doi: 10.1145/3080202.
[21] A. Kiani, N. Ansari, and A. Khreishah, “Hierarchical Capacity Provisioning for Fog Computing,” IEEE/ACM Trans. Netw., vol. 27, no. 3, pp. 962–971, 2019, doi: 10.1109/TNET.2019.2906638.
[22] S. R. Shishira, A. Kandasamy, and K. Chandrasekaran, “Workload Characterization: Survey of Current Approaches and Research Challenges,” in Proceedings of the 7th International Conference on Computer and Communication Technology, 2017, pp. 151–156, doi: 10.1145/3154979.3155003.
[23] M. Ciavotta, E. Gianniti, and D. Ardagna, “Capacity Allocation for Big Data Applications in the Cloud,” in Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion, 2017, pp. 175–176, doi: 10.1145/3053600.3053630.
[24] J. C. Mogul, R. Isaacs, and B. Welch, “Thinking about Availability in Large Service Infrastructures,” in Proceedings of the 16th Workshop on Hot Topics in Operating Systems, 2017, pp. 12–17, doi: 10.1145/3102980.3102983.
[25] J. Ericson, M. Mohammadian, and F. Santana, “Analysis of Performance Variability in Public Cloud Computing,” in 2017 IEEE International Conference on Information Reuse and Integration (IRI), 2017, pp. 308–314.
[26] I. Stypsanelli, O. Brun, S. Medjiah, and B. J. Prabhu, “Capacity Planning of Fog Computing Infrastructures under Probabilistic Delay Guarantees,” in 2019 IEEE International Conference on Fog Computing (ICFC), 2019, pp. 185–194.
[27] M. Torquato, L. Torquato, P. Maciel, and M. Vieira, “IaaS Cloud Availability Planning using Models and Genetic Algorithms,” in 2019 9th Latin-American Symposium on Dependable Computing (LADC), 2019, pp. 1–10.
[28] O. Biran et al., “Heterogeneous Resource Reservation,” in 2018 IEEE International Conference on Cloud Engineering (IC2E), 2018, pp. 141–147.
[29] L. Tang and H. Chen, “Joint Pricing and Capacity Planning in the IaaS Cloud Market,” IEEE Trans. Cloud Comput., vol. 5, no. 1, pp. 57–70, 2017.
[30] R. Vaze, “Online Knapsack Problem Under Expected Capacity Constraint,” in IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, 2018, pp. 2159–2167.
[31] B. Xia, T. Li, Q. Zhou, Q. Li, and H. Zhang, “An Effective Classification-based Framework for Predicting Cloud Capacity Demand in Cloud Services,” IEEE Trans. Serv. Comput., p. 1, 2018.
[32] C. Melo, R. Matos, J. Dantas, and P. Maciel, “Capacity-Oriented Availability Model for Resources Estimation on Private Cloud Infrastructure,” in 2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC), 2017, pp. 255–260.
[33] M. Noreikis, Y. Xiao, and A. Ylä-Jaäiski, “QoS-oriented capacity planning for edge computing,” in 2017 IEEE International Conference on Communications (ICC), 2017, pp. 1–6.
[34] K. N. Kumar and R. Mitra, “Resource Allocation for Heterogeneous Cloud Computing Using Weighted Fair-Share Queues,” in 2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), 2018, pp. 31–38.
[35] T. P. Roseline, C. J. M. Tauro, and M. Miranda, “An approach for efficient capacity management in a cloud,” in 2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), 2017, pp. 1–6.
[36] K. M. Maiyama, D. Kouvatsos, B. Mohammed, M.

 ISBN: 978-1-61208-827-3

Kiran, and M. A. Kamala, “Performance Modelling and Analysis of an OpenStack IaaS Cloud Computing Platform,” in 2017 IEEE 5th International Conference on Future Internet of Things and Cloud (FiCloud), 2017, pp. 198–205.
[37] H. A. Kholidy, “An Intelligent Swarm Based Prediction Approach For Predicting Cloud Computing User Resource Needs,” Comput. Commun., vol. 151, pp. 133–144, 2020, doi: https://doi.org/10.1016/j.comcom.2019.12.028.
[38] M. Liaqat et al., “Federated cloud resource management: Review and discussion,” J. Netw. Comput. Appl., vol. 77, pp. 87–105, 2017, doi: https://doi.org/10.1016/j.jnca.2016.10.008.
[39] V. K. Prasad, M. Shah, N. Patel, and M. Bhavsar, “Inspection of Trust Based Cloud Using Security and Capacity Management at an IaaS Level,” Procedia Comput. Sci., vol. 132, pp. 1280–1289, 2018, doi: https://doi.org/10.1016/j.procs.2018.05.044.
[40] N. Sadashiv, S. M. Dilip Kumar, and R. S. Goudar, “Cloud capacity planning and HSI based optimal resource provisioning,” in 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2017, pp. 1–6.
[41] M. Noreikis, Y. Xiao, and Y. Jiang, “Edge Capacity Planning for Real Time Compute-Intensive Applications,” in 2019 IEEE International Conference on Fog Computing (ICFC), 2019, pp. 175–184.
[42] S. Gupta and D. A. Dinesh, “Online adaptation models for resource usage prediction in cloud network,” in 2017 Twenty-third National Conference on Communications (NCC), 2017, pp. 1–6.
[43] C. Verbowski, E. Thayer, P. Costa, H. Leather, and B. Franke, “Right-Sizing Server Capacity Headroom for Global Online Services,” in 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), 2018, pp. 645–659.
[44] C. H. G. Ferreira et al., “A Low Cost Workload Generation Approach through the Cloud for Capacity Planning in Service-Oriented Systems,” 2017, doi: 10.1145/3018896.3018900.
[45] D. Ardagna et al., “Performance Prediction of Cloud-Based Big Data Applications,” in Proceedings of the 2018 ACM/SPEC International Conference on Performance Engineering, 2018, pp. 192–199, doi: 10.1145/3184407.3184420.
[46] J. C. Christopher, “Analytics Environments on Demand: Providing Interactive and Scalable Research Computing with Windows,” 2017, doi: 10.1145/3093338.3093369.
[47] M. Marin, V. Gil-Costa, A. Inostrosa-Psijas, and C. Bonacic, “Hybrid capacity planning methodology for web search engines,” Simul. Model. Pract. Theory, vol. 93, pp. 148–163, 2019, doi: https://doi.org/10.1016/j.simpat.2018.09.016.
[48] A. Brunnert and H. Krcmar, “Continuous performance evaluation and capacity planning using resource profiles for enterprise applications,” J. Syst. Softw., vol. 123, pp. 239–262, 2017, doi: https://doi.org/10.1016/j.jss.2015.08.030.
[49] T. Le Duc, R. G. Leiva, P. Casari, and P.-O. Östberg, “Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey,” ACM Comput. Surv., vol. 52, no. 5, 2019, doi: 10.1145/3341145.
[50] S. K. Moghaddam, R. Buyya, and K. Ramamohanarao, “Performance-Aware Management of Cloud Resources: A Taxonomy and Future Directions,” ACM Comput. Surv., vol. 52, no. 4, 2019, doi: 10.1145/3337956.
[51] D. Irwin and B. Urgaonkar, “Research Challenges at the Intersection of Cloud Computing and Economics,” National Science Foundation, USA, 2018.
[52] P. Mitrevski, F. Mitrevski, and M. Gusev, “A Decade Time-Lapse of Cloud Performance and Dependability Modeling: Performability Evaluation Framework,” 2019, doi: 10.1145/3320326.3320400.
[53] R. Han et al., “Workload-Adaptive Configuration Tuning for Hierarchical Cloud Schedulers,” IEEE Trans. Parallel Distrib. Syst., vol. 30, no. 12, pp. 2879–2895, 2019.
[54] P. Östberg et al., “Reliable capacity provisioning for distributed cloud/edge/fog computing applications,” in 2017 European Conference on Networks and Communications (EuCNC), 2017, pp. 1–6.
[55] M. Carvalho, D. A. Menascé, and F. Brasileiro, “Capacity planning for IaaS cloud providers offering multiple service classes,” Futur. Gener. Comput. Syst., vol. 77, pp. 97–111, 2017, doi: https://doi.org/10.1016/j.future.2017.07.019.
[56] K. C. Anupama, R. Nagaraja, and M. Jaiganesh, “A Perspective view of Resource-based Capacity planning in Cloud computing,” in 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019, pp. 358–363.
[57] M. Carvalho et al., “Multi-Dimensional Admission Control and Capacity Planning for IaaS Clouds with Multiple Service Classes,” in Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2017, pp. 160–169, doi: 10.1109/CCGRID.2017.14.
[58] E. Zharikov, O. Rolik, and S. Telenyk, “An integrated approach to cloud data center resource management,” in 2017 4th International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S T), 2017, pp. 211–218.
[59] R. I. Cartwright and B. Gilmer, “The Infinite Capacity Media Machine,” SMPTE Motion Imaging J., vol. 128, no. 9, pp. 1–7, 2019.
[60] B. T. Sloss, S. Nukala, and V. Rau, “Metrics That Matter,” Queue, vol. 62, no. 4, p. 88, 2018, doi: 10.1145/3305263.3309571.
[61] Elsevier Mendeley Tool. Available at https://www.mendeley.com/

 ISBN: 978-1-61208-827-3