2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion)

A Systematic Mapping Study on Real-time Cloud Services

Jakob Danielsson, Nandinbaatar Tsog, Ashalatha Kunnappilly
Mälardalen University, Västerås, Sweden
{jakob.danielsson, nandinbaatar.tsog, ashalatha.kunnappilly}@mdh.se

***Abstract*—Cloud computing is relatively a new technique to host and use the services and applications from the internet. Although it offers a multitude of advantages like scalability, low operating cost, accessibility and maintainability, etc., they are often not utilized to the fullest due to the lack of timeliness property associated with the cloud. Cloud services are mainly designed to maximize throughput and utilization of resources and hence incorporating predictable execution time properties in to the cloud is arduous. Nevertheless, cloud still remains a highly attractive platform for hosting real-time applications and services owing to features like elasticity, multi-tenancy, ability to survive hardware failures, virtualization support and abstraction layer support which provides flexibility and portability. In order for real-time safety-critical applications to exploit the potential of cloud computing, it is essential to ensure the predictable real-time behavior of cloud services. In this paper, we perform a systematic mapping study on real-time cloud services to identify the current research directions and potential research gaps. Our study focuses on analyzing the current architectures and software techniques that are available at present to incorporate real-time property of the cloud services. We also aim at investigating the current challenges involved in realizing a predictable real-time behavior of cloud services.**

***Index Terms*—Cloud computing, real-time, safety-critical applications, systematic mapping study.**

I. Introduction

The recent years have witnessed a tremendous increase in the usage of cloud computing technologies by the industries, and this has been accounted due to the fact that cloud services can deliver high performance solutions, with high flexibility. Cloud computing services can typically be used in any system which requires computational power, e.g., autonomous vehicle systems and big data centers. The characteristics of a cloud based service are that most of its computational power is located in one specific location and when it receives a computational request from a node, it performs the necessary computation and then distributes the results to the requesting node. A cloud system uses three types of services - Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [1].

There is a tremendous potential for cloud computing services in the present scenario, but in some cases, cloud computing cannot be used, due to the lack of real-time compatibility. Systems such as airbag systems and braking systems often contain strict real-time constraints to avoid disastrous consequences. In order to avoid such catastrophic consequences, all tasks inside a system should be executed within a predictable time interval, thus creating a real-time task schedule. Since a cloud computing service contains many layers which are dependent on each other, implementing real-time systems using cloud services can be very complex, e.g., a real-time task at the top SaaS layer can be dependent on the lower layers to achieve real-time performance. In addition, there are also other major challenges in achieving a real-time cloud such as dealing with scheduling algorithms, non-support for a real-time clock as an internal reference, delay in provisioning resources and virtual machines, lack of predictability of execution of tasks, etc. In this paper, we present a systematic mapping study [2], [3] on "real-time cloud services´´ with the aim of analyzing the techniques for achieving predictable real-time behavior in cloud computing scenario and also identifying papers which have addressed the challenges listed above.

The rest of the paper is organized as follows: Section 2 describes the process followed for our systematic mapping study, including screening methods, classification scheme, and initial search results. In Section 3, we describe in detail our results and analysis generated from the mapping study. Section 4 concludes the paper and gives some directions for future work.

II. Mapping Study Process

A systematic mapping study aims at providing a deep overview of researches conducted in a particular field and is also directed towards structuring the research results in an efficient manner to identify research gaps [2]. In order to perform a systematic mapping study on "Real-time cloud services" , we have followed an approach detailed by Petersen et.al [2]. The study approach is divided into 5 steps: (i) the definition of research questions, (ii) conducting the search, (iii) screening the papers, (iv) key wording and data extraction, and (v) the mapping process. The steps are illustrated in Figure 1. In the following subsections, we detail each of these steps with respect to our research focus area, which is 'Real-time cloud services'.

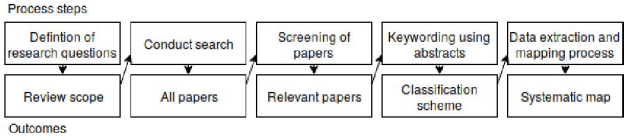

Figure 1: Systematic mapping process.

978-1-7281-0359-4/18/$31.00 ©2018 IEEE
DOI 10.1109/UCC-Companion.2018.00063

A. Research Questions

The first step of a mapping study comprises of framing the research questions to identify the scope of the research. The main goal of our study is to analyze how a predictable real-time behavior can be achieved within a cloud environment and to identify the hindrances, if any to achieve this behavior. This bigger goal is subdivided into 3 research questions to ease our analysis. The research questions are detailed below:

Research question 1*: *Which software techniques are most commonly used for achieving real-time behavior in a cloud environment?

The question aims at answering all software related queries and techniques, for instance, the type of scheduling algorithms and other software processes developed for achieving real-time behavior for cloud systems.

***Research question 2:** How is it possible to achieve real-time behaviour in the different cloud service layers?*

To answer this question, we explore how the existing cloud layers such as Saas, PaaS and IaaS are being used for supporting real-time applications and also identify some new frameworks, if any, that can support real-time behavior. Since there exist different layers for a cloud system, each one of them may have to be altered in order to achieve a predictable behavior of the cloud system.

***Research question 3:** What challenges can be identified for achieving a predictable real-time behavior in cloud?*

There are many challenges that are accounted for achieving predictable behavior even in the simplest real-time systems, such as scheduling algorithms and creating smarter cache designs. Therefore, the main challenges for ensuring predictability in complex systems such as cloud must not be left without investigation. There might also be some unknown challenges that expose themselves while we start experimenting with various real-time cloud solutions.

After the definition of research questions, we proceed to the Step 2 of our mapping study, which is conducting the search in various databases.

B. Conducting the search

In order to conduct the search, we first define a set of keywords that can be inserted into the search string for retrieving publications from various databases. In our study, we restrict the databases to IEEE, Scopus, Springer, ACM and ScienceDirect. We have not included Google Scholar results in our study as it retrieved us with a lot of non-peer reviewed articles The search queries we defined are based on our field of study, which is a combination of real-time and cloud. With the motive of getting all the papers related to real-time cloud services, we arranged our search query in such a way to contain either the key phrases “cloud computing” or “cloud services”, thus the full query we used is: “real-time” AND (“cloud computing” OR “cloud services”). Table I shows the search results generated.

Our initial search with the above search string extracted 3628 articles from IEEE, 255 from ACM, 11,221 from Scopus and 3754 from ScienceDirect. These numbers were huge to deal with and hence the works falling outside the research focus area should be eliminated and hence we proceed with the third step, screening the papers. We also made a search query on Springer, yielding 1,595,798, which potentially could have been due to a bug in the search query system as the search were conducted 2016. Using the search string on current writing date (October 2018), Springer yields 18,682 results in the year range from 2009 to 2016.

Database	Initial hits	By Title	By Abstract	Full text
IEEE	3628	70	34	23
ACM	255	14	12	7
Springer	1,595,798	121	12	9
Scopus	11,221	88	9	3
ScienceDirect	3,754	6	2	2

Table I: Search query results.

C. Screening the papers

As mentioned previously, a screening procedure is required to effectively select the relevant papers within the area of our study. The screening mechanism which we employ in our study is based on several inclusion-exclusion criteria. Our inclusion-exclusion criteria are explained in detail below.

1) **Inclusion Criteria:**
   a) Papers must be peer-reviewed.
   b) Papers should be based on techniques to achieve real-time cloud services, be it software related or architecture related.
   c) Papers must report challenges in incorporating timeliness property to cloud services.
   d) Papers must address the cloud from “cloud computing” and “cloud services” perspectives.
2) **Exclusion Criteria:**
   a) Papers must report on “cloud” related to “real-time”.
   b) All the non peer reviewed papers and articles in the form of abstracts, editorials or keynotes are excluded from the mapping study.
   c) Papers which are not in English language are excluded.
   d) Papers were selected in a timespan range of 2009-2016

The inclusion-exclusion criteria which we defined could successfully eliminate papers that were outside the scope of the current study. The screening procedure based on the above inclusion-exclusion criteria was conducted in 4 phases, including screenings of keywords, title, abstract and full text. At each phase, we uniformly divided the task at hand between all the authors and created a clear classification scheme.

1) ***Phase 0:*** In this phase, we have screened the papers based on the application of our search string to various database sources. The results of Phase 0 are already obtained in Step 2 and it yielded us with a total of 1,614,656 articles from different database sources. The search in different databases was distrubuted among the authors.

2) ***Phase 1:*** In Phase 1, we consider all the papers from Phase 0 and do a 'Title-based selection' and this could efficiently bring down the paper results when compared to Phase 0. By performing a title based search query, we could bring down the papers to 299 of which IEEE, ACM, Springer, Scopus and ScienceDirect constitute 70, 14, 121, 88 and 6 articles, respectively.
3) ***Phase 2:*** After screening the papers based on title, we proceeded to Phase 3, which is an abstract based selection. In this phase, we eliminated the duplicate papers and thus could bring down the paper count to 69 articles, of which IEEE has a contribution of 34, ACM 12, Springer 12, Scopus 3 and ScienceDirect 2 articles. All abstracts were read by each author.
4) ***Phase 3:*** In some of the papers, reading the abstract was insufficient to capture the necessary information and hence we proceeded by reading the full-text, mainly the introduction and conclusion sections. This could further narrow down our search to 44 papers, which we consider further for our mapping study. Each introduction and conclusion was also read by each author for categorizing the papers.

D. Key-wording technique and generating the classification scheme

The next step in our mapping study is to assign 'keywords' to each paper. Based on the research questions defined earlier, our motive was to identify the papers that were related to software techniques and architecture frameworks developed for achieving real-time cloud services. At the same time,

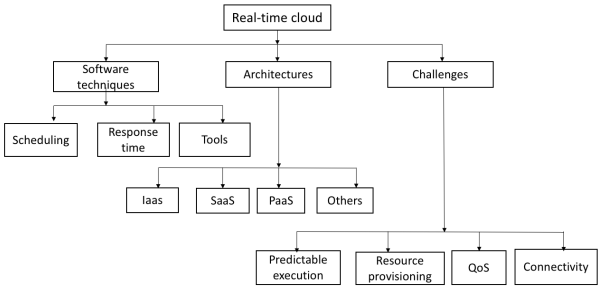

Figure 2: Classification Scheme.

we were also keen in analyzing the challenges for achieving predictable performance in cloud services. We found that reading the abstracts was not enough to generate efficient keywords and build a classification approach as there were domain overlaps and hence we proceeded by reading the introduction and conclusion as well and then utilized this information for key-wording. The keywords which we assigned are 'software-techniques', 'architecture-based approaches' and 'challenges'. After analyzing the contribution and context of research from the selected papers, a higher level view of the research was identified, which helped us to generate a more detailed classification scheme. The classification scheme is illustrated in Figure 2, the classification was also used for division of work, where one author read the full-text papers related to one category.

Publication type	2009	2010	2011	2012	2013	2014	2015	2016	Total	%
Conference	-	3	4	3	5	6	1	2	24	55
Journal	-	-	1	-	-	3	4	1	9	20
Book	-	1	-	-	-	2	1	2	6	14
Workshop	1	-	-	1	-	-	2	1	5	11
Total	1	4	5	4	5	12	8	6	44	100

Table II: Paper distribution by publication type.

The software techniques for achieving real-time cloud services were again categorized as scheduling based, response-time based and tool-based. Under scheduling based papers, we had papers discussing task scheduling, resource allocation and minimizing the total cost. Apart from that, there were also some papers dealing with the response-time analysis and some others reasoning about various methodologies/tools for achieving predictable cloud services, e.g, an improved cache based system and therefore we could classify them into categories named response- time based and tool-based papers. Under the category of architecture frameworks for achieving real-time services, we accomplished a detailed classification based on the major cloud platforms used like Saas, PaaS, and IaaS. The remaining real-time cloud architecture categories were classified into the category 'others', which included Control as a Service (CaaS) architectures, community cloud and graph based architectures. Identifying the challenges in real-time cloud computing was the trickiest part of the lot as each of the papers we analyzed presented its own limitations. However, we restricted our focus on identifying major challenges in achieving a predictable real-time environment and hence the challenges were further classified into papers mainly describing the hurdles to achieve predictable execution in cloud computing environment, challenges in successful resource provisioning for achieving real-time cloud services and also papers describing Quality of Service (QoS) and connectivity hindrances. The results are presented in detail in the next section.

III. Mapping Study Results

The previous sections unveiled the detailed mapping study approach which we followed for structuring the research results in the field of "Real-time cloud". In this section, we will discuss the results obtained by this mapping study. The results are categorized into 3 sub-sections. In the first subsection, we categorize our primary results based on research focus area, publication type and contribution type and also by the detailed classification scheme which we explained earlier. In the second subsection, we present our extensive mapping study results based on 44 papers we selected for performing the mapping study. We did a comparison-based approach by detailing the research contributions, approaches, scope of the works, limitations and results [3]. Finally, in the third subsection, we discuss the research gaps identified in our study.

Focus area	2009	2010	2011	2012	2013	2014	2015	2016	Total	%
Scheduling	-	2	2	-	2	6	1	3	16	36
Response time	-	-	-	-	1	-	1	-	2	5
Tools	-	1	1	3	-	-	1	-	6	14
IaaS	-	-	-	-	1	4	1	1	7	16
PaaS	-	1	-	1	1	-	-	-	3	7
SaaS	-	-	1	-	-	-	-	-	1	2
Others	-	-	-	-	-	-	2	-	2	5
Execution	-	-	1	-	-	1	1	-	3	7
Resources	1	-	-	-	-	-	-	1	2	5
QoS	-	-	-	-	-	-	1	-	1	2
Connectivity	-	-	-	-	-	-	-	1	1	2
Total	1	4	5	4	5	12	8	6	44	100

Table III: Paper distribution by focus area.

Contribution type	2009	2010	2011	2012	2013	2014	2015	2016	Total	%
Model	1	1	2	1	2	3	4	1	15	29
Process	1	-	-	-	-	1	-	-	2	4
Method	-	1	2	1	2	5	1	3	15	29
Tool	-	1	1	-	-	1	1	1	5	10
Concept-based	-	-	-	1	-	-	-	-	1	2
Survey	-	-	-	-	-	1	-	-	1	2
Metric	-	-	-	-	-	-	1	-	1	2
Architecture	-	1	1	-	1	1	-	1	5	10
Implementation	-	-	1	1	1	1	1	1	6	13
Total	2	4	7	4	6	14	8	7	51	100

Table IV: Distribution of primary studies by contribution type.

In this subsection, we illustrate the primary results from our systematic mapping study, generated from 44 articles, which comprised of various journal articles, conference and workshop proceedings and book chapters spanning across the years from 2009 to 2016. We have developed detailed classification schemes for the distribution of our primary studies based on publication type, research focus area and contribution type. The distribution of papers by publication type demonstrated that out of 44 papers we selected, 24 papers were conference proceedings, 9 were journal articles, 6 were book chapters and 5 were workshop proceedings. The results are tabulated in Table II. We performed another classification, aimed at finding out whether the paper falls under the category of software techniques, architectures or challenges tackled; 54% of the papers were software techniques related, 30% were architecture related and another 16% addressing the challenges. Further, we proceeded to identify the detailed division of focus areas based on the classification scheme which we detailed earlier, which are tabulated in Table III . By analyzing the results, we could identify that papers on software techniques for achieving real-time cloud were further classified as papers based on scheduling (16 papers, 36%), based on response time analysis (2 papers, 5%) and on tools (6 papers, 14%). Architecture-based papers are also subclassified as IaaS (7 papers, 16 %), PaaS (3 papers,7 %), SaaS (1 paper,2 %), and other categories (2 papers, 5%). And also, the papers reporting the challenges were also categorized into papers reporting on predictable execution challenges (3 papers,7%), resource provisioning challenges (2 papers, 5%), QoS issues (1 paper, 2%) and connectivity problems (1 paper, 2%).

We have also created a grouping of papers based on contribution type. The results are detailed in Table IV and show a distribution of 29% of papers based on model (15 papers), 4% on process (2 papers), 29% on methodology (15 papers), 10% on tool (5 papers), 2% concept-based (1 paper), 2% survey-based (1 paper), 2% metric-based (1 paper), 10% based on architecture (5 papers) and another 13% based on implementation (6 papers).

A. Mapping study results

We have done a deeper analysis on 44 papers which we have selected for the mapping study. Our approach consisted of identifying the major contributions of the papers and then classifying it based on our classification scheme, the scope of research and approaches. We also proceeded to analyze the results of the articles and then checked whether the results are experimentally validated or not. In each of the paper, we also tried to identify the limitations or constraints encountered in following a particular methodology. This approach helped us to efficiently structure the research results within the area of real-time cloud computing and assisted us further in identifying the research gaps in the area. The research gaps identified will definitely pave the way for future researches in this area. The major research gaps which we identified are discussed in the next section. Due to space constraints, we discuss the results briefly here, however the extensive mapping study results, including the review of each paper with a focus on the problem addressed, the scope of the approach, limitations/constraints and results of the validation of the approach can be accessed online [1].

B. RQ1 - Software techniques for real-time systems

We found 22 papers addressing software techniques for real-time systems of which many point to the problem associated with real-time task scheduling while still maintaining energy efficiency. Various techniques such as combining Earliest Deadline First (EDF) with First Come First Served (FCFS) [4], using a Paused Rate Monotonic (PRM) scheduler [5], and improving EDF by adding laxity as a parameter to the algorithm [6] are examples that demonstrate novel ways to improve real-time performance of cloud systems by altering existing algorithms. In another interesting work, the authors suggest creating new scheduling algorithms such as the two-tier cooperative task scheduling approach [7], working on preemptive online scheduling for tasks [8], allocating resources for tasks based on profit and penalty [9] and particle swarm optimized scheduling [10]. Finally, Chen et al. [11] address the issue of uncertainty in a computational environment by implementing an algorithm named Proactive and Reactive Scheduling (PRS) which handles real-time tasks and outperforms other related algorithms. In contrast, Huynh et al. [12] implement the Cost-Efficient Real-time Application Scheduling (CERAS) algorithm which creates a schedule according to the trade-off between response time and cost efficiency.

[1] https://github.com/hylz/Mapping

Scheduling also enables efficient resource allocation. Yuhan Du and Gustavo De Veciana [13] show that it is possible to make substantial resource savings by prioritizing tasks which have the largest QoS deficits. Tsai et al. [14] propose a model on how to partition databases to achieve full isolation of a system. In addition, Mora et al. [15] developed a framework for using remote computing resources to meet real-time systems constraints and it proved to be a simple way of creating task schedules with remote resources. Chawarut and Woraphon [16] present a CPU re-allocation strategy that improves energy management and execution time of tasks. Finally, Hoffert et al. [17] investigate on how to apply autonomous configuration of a fog environment using the ADAptive Middleware And Network Transports (ADAMANT) framework, which is based on the available computing resources. Virtualization for cloud services is also a common theme among these references, as it often provides some level of task isolation. Wu et al. [18] proposed a mechanism for the Xen hypervisor to provide better guarantees to real-time constraints in cloud-based systems. In another study, Lundberg and Shirinbab [19] analyze how current real-time theory can be applied to cloud services in a virtualized environment. Other work include investigating the energy efficiency [20] of virtualized environments and those focusing on minimizing the energy consumption [21] while maintaining a probabilistic behavior. However, using a virtualized environment often comes with migration overhead. Zhang et al. [22] worked on this aspect and propose a scheduling algorithm which outperforms other algorithms in terms of energy efficiency while maintaining deadline guarantees without migration overhead. Understanding the resource usage in a cloud environment is a critical aspect, as it enables optimization of previously unknown bottlenecks in a system. Alhamazani et al. [23] implemented a cross-layer benchmark for cloud environments which can be used for ensuring runtime QoS. Kyongo et al. [24] extended the benchmarking by monitoring virtual resources to improve real-time properties such as jitter, latency and scalability, which was shown to outperform the previous "REST-ful´´ monitoring approach. Resource allocation can also be critical to real-time cloud systems, as shown by Kumar et al. [25] whom create a novel method for allocating resources efficiently. Other papers related to RQ1 addresses real-world scenarios where real-time cloud applications are necessary. Krishnappa et al. [26] finds the ExoGENI algorithm to be most feasible for handling large scale weather forecasting data sets in real-time. Lin et al. [27] show how it is possible to create real-time carpool services using a mobile client and a global cloud carpool system. Finally, Esen et al. [28] investigate the Control as a Service model for handling real-time constraints in autonomous cars.

C. RQ2 - Cloud architectures specific to real-time systems

The IaaS layer uses the hardware of cloud machines, and hence it can become tricky to guarantee real-time constraints due to the heterogeneity of a chip as well as resource contention of multiple cores. Cordeschi et al. [29] discuss and validate that hard-real time capabilities of cloud systems using a virtual machine that is achievable through a dynamic trade-off. Real-time systems may also need fault tolerance, hence, Kumar et al. [30] implement the High Adaptive Fault Tolerance in Real time Cloud computing (HAFTRC) model that successfully makes a more reliable system by selecting the most reliable virtual machine according to the reliability of the Random Access Memory (RAM), Million Instructions Per Second (MIPS), bandwidth, cloudlets, and more. Mohammed et al. [31] take another approach and optimize the success rates of virtual nodes and machines ultimately leading to faults being repaired before deadline misses occur. Hypervisors make the core of a virtualized environment, but cannot be run with real-time compatibility and Xi et al. [32] developed the RT-openstack to co-host real-time- and regular operating systems. Furthermore, Denizak and Bak [33] investigate how cloud can be used in real-time and how it can be used to satisfy all user needs by implementing an iterative algorithm on a distributed architecture. Finally, Dutra Ös and Bressan [34] proposed an IaaS based community cloud architecture that ensures real-time properties through scheduling.

Managing multimedia content often include work intense calculations over large-scale data-sets. Therefore, cloud computing can often be applied to increase the performance of such applications while potentially maintaining real-time performance. Boniface et al. [35] suggest a PaaS architecture model to handle scenarios such as augmented reality while providing which include metrics such as QoS specification, event prediction, and dynamic Service Level Agreement (SLA) navigation to ensure real-time execution. Von Söhsten and Murilo [36] suggested an approach combining windows Azure and Emgu CV, proving it is possible to effectively utilize cloud-services for face recognition. Wang et al. [37] investigated on creating real-time networks for vehicular systems using a three-tier V-cloud architecture to achieve real-time performance for systems that include actuator units, communication media and server media and Piyare et al. [38] suggest a platform to easily integrate wireless sensor networks in to the cloud. Belli et al. [39] present a graph-based cloud architecture [39] which is shown to produce a decreased delay in real-time streams.

D. RQ3 - Challenges for real-time in cloud environments

Our last research question focuses on finding papers related to future challenges for achieving real-time capabilities in cloud environment. One major challenge which has not yet been completely solved is the shared internal memory for multiple cores. Xu et al. [40] present a cache-aware compositional analysis technique, used for timing analysis of components scheduled on a multi-core platform. The results show a significantly improved resource bandwidth usage as well as a reduced cache miss ratio. With similar goals, Zhang et al. [41] propose a distributed layered cache hierarchy built on the Hadoop Distributed file system for real-time cloud services, maintaining a hit-ratio of 95%. Hoon et al. [42], [20] take one leap ahead and focus on the virtual machine domain, where a real-time cloud service framework is suggested for requesting virtual platforms, which is validated by evaluating

simulations of power-aware real-time services. Other domains include health-care, where the data from patients in the critical care unit needs to be handled in real-time. McGregor [43] investigates this issue by using the Artemis cloudsuite to enable multidimensional real-time analysis of data.

Other relevant papers consider every-day scenarios such as online shooter gaming [44], Netflix streaming [45] and real-time collaborative editing [46] where a non-deterministic timing behavior may not cause great harm but nevertheless cause user experience to falter.

Moreover, García-Valls et al. [47] performed a similar mapping study on identifying technical challenges in supporting real-time applications in cloud technologies. The authors conclude it is hard to achieve time-predictability within virtualized cloud systems due to the limited access to the hardware. This is the only mapping mapping study that we could retrieve with our search string. García-Valls et al. was the only mapping study found using our defined search query.

E. Discussion

In this section, we detail about the potential research gaps that we have identified based on our mapping study.

1) Lack of unified methods and integrated architecture support for achieving real-time cloud services: The existing research work in the area of real-time cloud architectures and methodologies lacks a unified approach or architecture which can be used to reliably host real-time applications. Though most of the existing work is aimed at altering the existing IaaS platform to suit the real-time functionality, we could not find an approach integrating the various platforms like IaaS, PaaS, SaaS, which we believe is very essential to further exploit the advantages of real-time cloud.

2) Problems hindering the predictable execution of tasks in cloud platform: Cloud platforms offer a lot of challenges while integrating it with real-time computing environment. Practically, all real-time safety critical applications demand high performance computing where performance sustainability, resource guarantees, and timely guarantee for results are highly essential, and cloud environment is simply not designed to handle all of these. Therefore, the use of cloud for hosting hard real-time applications is still beyond realities. These issues become more troublesome in a multi-core environment, where it is very complex to handle virtualization and multicore timing.

In addition to these issues, cloud has a sub-optimal physical topology intended to accommodate a large number of applications, which gives rise to performance penalties, that cannot be accommodated in real-time computing. It is also complex to account for the delays in the provisioning of resources and virtual machines to ensure predictability. Although some work has proceeded in this direction, they need to mature a lot to use it for practical applications.

3) Lack of efficient methods for identifying QoS based degradation in real-time cloud services: Yet another important concern is maintaining the desired QoS for real-time cloud services. There are a lot of QoS issues that need to be handled in a real-time scenario. For e.g, ensuring fairness when a service goes down, especially when you use it as a pay as you go manner, is very difficult and is a serious problem for all soft real-time applications in cloud. Moreover, achieving scalability of real-time online interactions in cloud is a complex concept. There are also many other issues, e.g., effectively managing power in the systems in data-centers while they are used for real-time applications, improving the performance of file access in order to simultaneously retrieve a large amount of data for achieving real-time cloud services etc., which need to be investigated beyond the scopes of the present literature, before which the potential of cloud computing cannot be fully utilized to host real-time applications.

IV. Conclusion

In this paper, we have conducted a systematic mapping study of real-time cloud services to structure the research results and establish detailed state-of-the-art research effectively. Our mapping study could identify some potential research gaps in the existing literature. The major issue which we found is the considerable lack of efficient works in this area that can serve as a base for future researches. Almost all of the existing works are fragmented, with no support for an integrated architecture for achieving predictability of cloud services. We earnestly hope that the future works in this area will definitely proceed in these directions and address the existing challenges.

References

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica *et al.*, "Above the clouds: A berkeley view of cloud computing." Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, 2009.

[2] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *12th international conference on evaluation and assessment in software engineering*, vol. 17, no. 1. sn, 2008.

[3] A. Abdelmaboud, D. N. Jawawi, I. Ghani, A. Elsafi, and B. Kitchenham, "Quality of service approaches in cloud computing: A systematic mapping study," in *The Journal of Systems and Software 101*, 2015, p. 159–179.

[4] S. P. Reddy and H. Chandan, "Energy aware scheduling of real-time and non real-time tasks on cloud processors (green cloud computing)," in *Information Communication and Embedded Systems (ICICES), 2014 International Conference on*. IEEE, 2014, pp. 1–5.

[5] F. Teng, F. Magoulès, L. Yu, and T. Li, "A novel real-time scheduling algorithm and performance analysis of a mapreduce-based cloud," *The Journal of Supercomputing*, vol. 69, no. 2, pp. 739–765, 2014.

[6] T.-Y. Chen, H.-W. Wei, J.-S. Leu, and W.-K. Shih, "Edzl scheduling for large-scale cyber service on real-time cloud," in *2011 IEEE International Conference on Service-Oriented Computing and Applications (SOCA)*. IEEE, 2011, pp. 1–3.

[7] S. Hosseinimotlagh, F. Khunjush, and S. Hosseinimotlagh, "A cooperative two-tier energy-aware scheduling for real-time tasks in computing clouds," in *2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*. IEEE, 2014, pp. 178–182.

[8] R. Santhosh and T. Ravichandran, "Pre-emptive scheduling of on-line real time services with task migration for cloud computing," in *Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on*. IEEE, 2013, pp. 271–276.

[9] S. Liu, G. Quan, and S. Ren, "On-line scheduling of real-time services for cloud computing," in *2010 6th World Congress on Services*. IEEE, 2010, pp. 459–464.

[10] H. Chen and W. Guo, "Real-time task scheduling algorithm for cloud computing based on particle swarm optimization," in *International Conference on Cloud Computing and Big Data in Asia*. Springer, 2015, pp. 141–152.

[11] H. Chen, X. Zhu, H. Guo, J. Zhu, X. Qin, and J. Wu, "Towards energy-efficient scheduling for real-time tasks under uncertain cloud computing environment," *Journal of Systems and Software*, vol. 99, pp. 20–35, 2015.

[12] C.-T. Huynh, T.-D. Nguyen, H.-Q. Nguyen, and E.-N. Huh, "Cost efficient real-time applications scheduling in mobile cloud computing," in *Proceedings of the Fifth Symposium on Information and Communication Technology*. ACM, 2014, pp. 248–255.

[13] Y. Du and G. de Veciana, "Scheduling for cloud-based computing systems to support soft real-time applications," *arXiv preprint arXiv:1601.06333*, 2016.

[14] W. T. Tsai, Q. Shao, X. Sun, and J. Elston, "Real-time service-oriented cloud computing," in *2010 6th World Congress on Services*, July 2010, pp. 473–478.

[15] H. Mora Mora, D. Gil, J. F. Colom López, and M. T. Signes Pont, "Flexible framework for real-time embedded systems based on mobile cloud computing paradigm," *Mobile Information Systems*, vol. 2015, 2015.

[16] W. Chawarut and L. Woraphon, "Energy-aware and real-time service management in cloud computing," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2013 10th International Conference on*. IEEE, 2013, pp. 1–5.

[17] J. Hoffert, D. C. Schmidt, and A. Gokhale, "Adapting distributed real-time and embedded pub/sub middleware for cloud computing environments," in *Proceedings of the ACM/IFIP/USENIX 11th International Conference on Middleware*. Springer-Verlag, 2010, pp. 21–41.

[18] S. Wu, L. Zhou, D. Fu, H. Jin, and X. Shi, "A real-time scheduling framework based on multi-core dynamic partitioning in virtualized environment," in *IFIP International Conference on Network and Parallel Computing*. Springer, 2014, pp. 195–207.

[19] L. Lundberg and S. Shirinbab, "Real-time scheduling in cloud-based virtualized software systems," in *Proceedings of the Second Nordic Symposium on Cloud Computing & Internet Technologies*. ACM, 2013, pp. 54–58.

[20] K. H. Kim, A. Beloglazov, and R. Buyya, "Power-aware provisioning of virtual machines for real-time cloud services," *Concurrency and Computation: Practice and Experience*, vol. 23, no. 13, pp. 1491–1505, 2011.

[21] S. Hosseinimotlagh, F. Khunjush, and R. Samadzadeh, "Seats: smart energy-aware task scheduling in real-time cloud computing," *The Journal of Supercomputing*, vol. 71, no. 1, pp. 45–66, 2015.

[22] Y. Zhang, L. Chen, H. Shen, and X. Cheng, "An energy-efficient task scheduling heuristic algorithm without virtual machine migration in real-time cloud environments," in *International Conference on Network and System Security*. Springer, 2016, pp. 80–97.

[23] K. Alhamazani, R. Ranjan, P. P. Jayaraman, K. Mitra, F. Rabhi, D. Georgakopoulos, and L. Wang, "Cross-layer multi-cloud real-time application qos monitoring and benchmarking as-a-service framework," 2015.

[24] K. An, S. Pradhan, F. Caglar, and A. Gokhale, "A publish/subscribe middleware for dependable and real-time resource monitoring in the cloud," in *Proceedings of the Workshop on Secure and Dependable Middleware for Cloud Monitoring and Management*. ACM, 2012, p. 3.

[25] K. Kumar, J. Feng, Y. Nimmagadda, and Y.-H. Lu, "Resource allocation for real-time tasks using cloud computing," in *Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on*. IEEE, 2011, pp. 1–7.

[26] D. K. Krishnappa, E. Lyons, D. Irwin, and M. Zink, "Network capabilities of cloud services for a real time scientific application," in *Local Computer Networks (LCN), 2012 IEEE 37th Conference on*, Oct 2012, pp. 487–495.

[27] C.-H. Lin, M.-K. Jiau, and S.-C. Huang, "A cloud computing framework for real-time carpooling services," in *Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in*, Oct 2012, pp. 266–271.

[28] H. Esen, M. Adachi, D. Bernardini, A. Bemporad, D. Rost, and J. Knodel, "Control as a service (caas): cloud-based software architecture for automotive control applications," in *Proceedings of the Second International Workshop on the Swarm at the Edge of the Cloud*. ACM, 2015, pp. 13–18.

[29] N. Cordeschi, D. Amendola, F. D. Rango, and E. Baccarelli, "Networking-computing resource allocation for hard real-time green cloud applications," in *2014 IFIP Wireless Days (WD)*, Nov 2014, pp. 1–4.

[30] P. Kumar, G. Raj, and A. K. Rai, "A novel high adaptive fault tolerance model in real time cloud computing," in *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference* -, Sept 2014, pp. 138–143.

[31] B. Mohammed, M. Kiran, I. U. Awan, and K. M. Maiyama, "Optimising fault tolerance in real-time cloud computing iaas environment," in *2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud)*, Aug 2016, pp. 363–370.

[32] S. Xi, C. Li, C. Lu, C. D. Gill, M. Xu, L. T. X. Phan, I. Lee, and O. Sokolsky, "Rt-open stack: Cpu resource management for real-time cloud computing," in *2015 IEEE 8th International Conference on Cloud Computing*, June 2015, pp. 179–186.

[33] S. Deniziak and S. Bak, "Synthesis of real time distributed applications for cloud computing," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*. IEEE, 2014, pp. 743–752.

[34] M. D. Ös and G. Bressan, "A community cloud for a real-time financial application-requirements, architecture and mechanisms," in *International Conference on Algorithms and Architectures for Parallel Processing*. Springer, 2014, pp. 364–377.

[35] M. Boniface, B. Nasser, J. Papay, S. C. Phillips, A. Servin, X. Yang, Z. Zlatev, S. V. Gogouvitis, G. Katsaros, K. Konstanteli *et al.*, "Platform-as-a-service architecture for real-time quality of service management in clouds," in *Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on*. IEEE, 2010, pp. 155–160.

[36] D. von Söhsten and S. Murilo, "Multiple face recognition in real-time using cloud computing, emgu cv and windows azure," in *2013 13th International Conference on Intellient Systems Design and Applications*, Dec 2013, pp. 137–140.

[37] J. Wang, J. Cho, S. Lee, and T. Ma, "Real time services for future cloud computing enabled vehicle networks," in *Wireless Communications and Signal Processing (WCSP), 2011 International Conference on*, Nov 2011, pp. 1–5.

[38] R. Piyare, S. Park, S. Y. Maeng, S. H. Park, S. C. Oh, S. G. Choi, H. S. Choi, and S. R. Lee, "Integrating wireless sensor network into cloud services for real-time data collection," in *2013 International Conference on ICT Convergence (ICTC)*, Oct 2013, pp. 752–756.

[39] L. Belli, S. Cirani, G. Ferrari, L. Melegari, and M. Picone, "A graph-based cloud architecture for big stream real-time applications in the internet of things," in *European Conference on Service-Oriented and Cloud Computing*. Springer, 2014, pp. 91–105.

[40] M. Xu, L. T. X. Phan, O. Sokolsky, S. Xi, C. Lu, C. Gill, and I. Lee, "Cache-aware compositional analysis of real-time multicore virtualization platforms," *Real-Time Systems*, vol. 51, no. 6, pp. 675–723, 2015.

[41] J. Zhang, Q. Li, and W. Zhou, "Hdcache: A distributed cache system for real-time cloud services," *Journal of Grid Computing*, pp. 1–22, 2016.

[42] K. H. Kim, A. Beloglazov, and R. Buyya, "Power-aware provisioning of cloud resources for real-time services," in *Proceedings of the 7th International Workshop on Middleware for Grids, Clouds and e-Science*. ACM, 2009, p. 1.

[43] C. McGregor, "A cloud computing framework for real-time rural and remote service of critical care," in *Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on*, June 2011, pp. 1–6.

[44] D. Meiländer and S. Gorlatch, "Modelling the scalability of real-time online interactive applications on clouds," in *Proceedings of the Third International Workshop on Adaptive Resource Management and Scheduling for Cloud Computing*. ACM, 2016, pp. 14–20.

[45] M. Aazam and E.-N. Huh, "Qos degradation based reimbursement for real-time cloud communication," in *Proceedings of the 1st Workshop on All-Web Real-Time Systems*. ACM, 2015, p. 6.

[46] T. Ozono, R. M. E. Swezey, S. Shiramatsu, T. Shintani, R. Inoue, Y. Kato, and T. Goda, "A real-time collaborative web page editing system wfe-s based on cloud computing environment," in *Advanced Applied Informatics (IIAIAAI), 2012 IIAI International Conference on*, Sept 2012, pp. 224–229.

[47] M. García-Valls, T. Cucinotta, and C. Lu, "Challenges in real-time virtualization and predictable cloud computing," *Journal of Systems Architecture*, vol. 60, no. 9, pp. 726–740, 2014.