Architectural Run-Time Models for Operator-in-the-Loop Adaptation of Cloud Applications

Robert Heinrich¹, Reiner Jung², Eric Schmieders³, Andreas Metzger², Wilhelm Hasselbring², Ralf Reussner¹, and Klaus Pohl³

¹Karlsruhe Institute of Technology, Germany, {heinrich, reussner}@kit.edu
²Kiel University, Germany, {reiner.jung, hasselbring}@email.uni-kiel.de
³Paluno, University Duisburg-Essen, Germany, {eric.schmieders, andreas.metzger, klaus.pohl}@paluno.uni-duke.de

Abstract—Building software systems by composing third-party cloud services promises many benefits. However, the increased complexity, heterogeneity, and limited observability of cloud services brings fully automatic adaptation to its limits. We propose architectural run-time models as a means for combining automatic and operator-in-the-loop adaptations of cloud services.

I. INTRODUCTION

Two major trends continue to significantly shape the way software engineers build and maintain future long-living software systems: (I) Software systems will be built by selecting, configuring, and composing third-party software-defined services, providing access to application software, cloud compute and storage facilities, internet-connected things, as well as data. Software-defined services separate ownership, maintenance and operation from usage of software. Service users do not need to deploy and run software on their own. They use software executed by third parties that can be remotely accessed through service interfaces [1]. (II) Software will be deployed on virtualized, remote compute infrastructures, such as cloud infrastructures, software-defined network devices, as well as Internet-of-Things devices. These deployment models result in software that will increasingly be deployed on hardware resources and on top of middleware that is owned and operated by third parties. The use of third party services and cloud infrastructures promises many benefits, such as flexibility, scalability, reusability and economic use of resources. Yet, the increased complexity and unprecedented heterogeneity of cloud services as well as their limited observability, will bring fully automatic adaptation to its limits. The distribution of cloud services across different legal systems, e.g., the EU and the USA, with different regulations for privacy and data proliferation, apply restrictions on the deployment and potential adaptations. The complexity of effects of external events lead to major challenges (e.g., see [2]). For instance, business decisions, such as sales campaigns, lead to an intensive application usage which cannot be anticipated by the system when monitoring data alone. Other effects, such as weather conditions and economic turmoil, may have an impact on customer behavior resulting in varying application utilization. Fully capturing and formalizing the knowledge required to take autonomic adaptation decisions may just become infeasible or may face prohibitive costs.

While, in recent years, the research focus for adaptation was on driving further its automation (e.g., see [3]), it becomes evident that we need to allow operators (humans) to engage in the adaptation process. An open question is how to facilitate such operator-in-the-loop adaptation. This paper takes a model-based stance by proposing architectural run-time models as a means for combining automatic and operator-in-the-loop adaptations of cloud services. A run-time model constitutes an abstraction of the software system of consideration (see [4]). Such models can be used as prescriptive models, visualizing the currently executing system. In addition, they may also be used as prescriptive models, specifying the envisioned future state of the software system.

In this paper we present first results towards supporting the automatic and operator-in-the-loop adaptation of cloud-based systems by leveraging architectural run-time models for planning and executing adaptations, while taking into account multiple quality criteria. Our envisioned solution complements the iObserve approach by extending it to the planning and execution phases of the MAPE loop (iObserve so far focused on the monitoring and analysis phases [5], [6]). A set of adaptation candidates is computed and in case where selection among those candidates is not feasible automatically, alternatives are presented to the operator for decision making. This facilitates an operator to engage with the adaptation process in three distinct areas. Firstly, by allowing introducing external workload scenarios which are tailored to fit predictions regarding a specific external event. Secondly, by providing the means to observe, evaluate, assess and manipulate scenario candidates visually. Thirdly, by supporting the operator in translating adaptation candidates to actual adaptation tasks.

The remainder of the paper is structured as follows. Sec. II introduces iObserve as a basis for this paper’s contributions. Sec. III discusses limitations of current adaptation mechanisms in the literature. Sec. IV sketches the main ideas and concepts for using architectural run-time models for operator-in-the-loop adaptation of cloud services. The paper concludes in Sec. V.

This work was partially supported by the DFG (German Research Foundation) under the Priority Programme SPP1593: Design For Future – Managed Software Evolution.
II. The iObserve Approach

iObserve [7] addresses the challenges of distributed cloud-based software systems by following the widely adopted MAPE control loop model for adaptation and extends it with shared models to ease the transition between adaptation and evolution. For adaptation, the state of the software system is determined by monitoring, analyzed to detect anomalies and predict deviations, used as input for planning the adaptation to mitigate anomalies, and finally to execute the adaptation. iObserve supports monitoring and analysis of cloud-based software systems through architectural run-time models [5] to facilitate the automated analysis, as well as the human inspection to detect, in particular, performance and privacy anomalies. In our work we focus on performance, privacy and costs as especially in the cloud context these aspects are closely interrelated. The application usage impacts on the application’s performance. Continuously appraised elasticity rules trigger the migration and replication of cloud application’s software components among geographically distributed data centers. Both, migration and replication, may increase performance yet lead to violation of privacy policies and increasing costs. The architectural run-time models are derived and updated or reverse engineered via dynamic analysis of monitoring data [8]. As an umbrella to integrate design-time models, code generation, monitoring, analysis, and run-time model update, iObserve introduced a concise megamodel [5] which is divided into four sections defined by two dimensions: one for design-time vs. run-time, and one for model vs. implementation level. At design-time the megamodel shows how an application model including an architectural model is combined with our model-driven monitoring approach [8] and used for code generation. On the run-time side, monitoring data which relates to implementation artifacts are associated with run-time application model elements based on a correspondence model [9]. The megamodel builds upon an aspect-oriented modeling approach [10]. Various analyses are performed on the monitoring data and used to update the application model. Further description of iObserve, background information and technical details are given in [5].

iObserve currently is focused on monitoring and analysis of cloud-based software systems. Addressing planning and execution phases of the MAPE control loop requires facing various challenges due to the complex and heterogeneous nature of cloud services. While different directions of further development are possible, we consider the following three pillars as next steps for evolving iObserve.

First, the distribution of cloud services across several legal systems with varying privacy regulations requires putting special emphasis on privacy-driven planning. Second, the complexity of cloud services and the complexity of effects of external events lead to major challenges in finding an appropriate system design. Therefore, we put special emphasis on design space exploration to address multiple quality criteria (esp. performance, privacy and cost) based on models to be explored for solutions. Third, restrictions in fully automatic adaptation need to allow human operators to engage in the adaptation process. Therefore, we put special emphasis on operator-in-the-loop adaptation to overcome limitations of automatic adaptation. For all three pillars architectural run-time models are key artifacts to address these challenges.

III. State of the Art

This section discusses existing work and its limitations in context of the three pillars of the envisioned approach.

Privacy-driven planning, as an adaptation activity during run-time, is an emerging field of research. It combines cloud characteristics such as elasticity and shared ownership. Existing research tackles the enforcement of privacy policy compliance from two different angles – privacy-by-design and privacy audits. Further, planning techniques have been developed for quality attributes such as privacy and reliability. Privacy-by-design employs two main concepts: access control and privacy-aware deployment and elasticity rules. Research on access control mechanisms (e.g. [11]) investigates how to equip components with mechanisms that permit or grant data access after matching the client characteristics with the data policies. However, access control mechanisms face difficulties in the cloud whenever the controlled component is migrated or replicated across data centers (e.g., a database). Elasticity and deployment rules are typically defined during design-time. However, rules that implement data geo-location policies have to be defined considering the data stored by a virtual machine as well as the data, which may be transferred to it. Yet, this information is not available during design-time, and thus rules defined during design-time may not accurately capture geo-location constraints at run-time.

Approaches for privacy audits of cloud services (e.g. [12]) use host geo-location and provable data possession to evaluate the compliance of privacy regulations. Host geo-location can be checked based on ping round trip times for a service interface utilizing different ping origins. This allows to decide whether a service interfaces reside at geo-locations excluded in privacy policies. However, the software components behind the service interfaces might be migrated or replicated, while the service interface remains at the same geo-location. Fundamentals for facilitating privacy-driven planning for cloud adaptations has been proposed in [13], [14]. Further, existing work focuses on the identification of privacy violations and does not cover planning or adaptation aspects.

Quality aware planning techniques employ the phases of the MAPE loop to adapt cloud applications as soon as, e.g., performance issues emerge. These techniques exploit the proportional relation between allocated hardware and performance. For instance, calibrated descriptive run-time models [15] are used in performance predictions to steer the planning activities and adjust the number of hardware nodes to reach the desired performance. However, resolving the root causes of privacy violations is inherently different from resolving performance issues. While adding additional hardware may increase system performance it will not, e.g., avoid personal data being moved to excluded geo-locations.
To summarize, existing work on privacy-driven-planning neglects two essential characteristics of cloud applications, (1) the dynamic migration and replication of cloud application components (cloud elasticity) across data centers [16]; (2) the limited control and visibility of cloud elasticity by cloud customers (shared ownership).

**Design space exploration:** Many cloud based applications serve thousands if not millions of users (e.g., 425 million users for gmail) whose behavior cannot be fully understood at design-time, because their behavior may change over time due to a wide variety of external and internal events, such as seasonal sales and campaigns. Furthermore, service offers and cloud products change over time in performance and cost. In combination with privacy concerns, this yields a complex problem space where multiple solutions can be computed forming a design space. Design space exploration requires several different input models describing various aspects of the design space, in particular user behavior, workload intensity, cloud profiles covering cost and resource descriptions, and privacy concerns. These models are then explored for solutions.

Workload characterization approaches used for adaptive systems and workload generation [17] focus on the construction of workload intensity pattern composed of statistical functions and recorded request sets. However, the intensity alone is not sufficient to understand the workload and predict future workloads. Therefore, a wide variety of methods, including clustering and Markov chain models of service utilization [18] are used to model user behavior and workloads [19]. These models can be derived from observation data [5] and must reflect the business impact on system usage [20]. However, the discrimination of different user behaviors and the aggregation of similar behaviors is complicated. The BEAR approach [21] addresses the construction and analysis of such behavior models. However, BEAR focuses on the engineer and the use of the analysis at design-time, e.g., in an evolution step. Furthermore, its user distinction is based on IP addresses which does not work well in the present condition of the Internet, especially for mobile devices, where many users share the same IP address based on network address translation.

Cloud profiles describe various properties of cloud services including resources and cost, like in the CloudMIG approach [22], [23]. While CloudMIG originates in the software migration domain, its cloud profiles can also be used in the adaptation context. Based on these models, design space exploration can be performed. An overview of design space exploration methods is given in [24]. Basically, research in this area can be distinguished into three different groups. Rule-based approaches can only change models according to predefined rules and are limited to single quality attributes. Specialized optimization approaches only consider predefined degrees of changes. Meta heuristic-based approaches use degrees of freedom in models but mostly do not enable trade-off decisions. CDOXplorer [25] employs genetic algorithms and simulation to assess the fitness of cloud development options based on cloud profiles.

To summarize, existing work on design space exploration neglects the differentiation of user behavior and trade-off decisions between different quality attributes.

**Operator-in-the-loop adaptation:** In the operation of software systems, the main goal was to automate adaptation and eliminating human involvement. To the best of our knowledge, the combination of automatic and operator-in-the-loop adaptation is not covered in related work. However, the increasing complexity of systems require to address this subject. Hence, we rely on other research areas as source for operator-in-the-loop approaches. Combined automatic and operator adaptation is addressed in so-called supervisory control and data acquisition (SCADA) systems [26] designed to control and operate large scale distributed industrial plants and infrastructure networks, such as gas, water, and power supply. In all these systems, operators are included in the control process as complete automation may result in undesirable states.

In the software engineering domain, recommendation systems [27] are used for software development and evolution, as well as, other design-time tasks. They use source code, version history and issue tracking systems, as corpus to derive knowledge from. For example, source code acquired from other projects utilizing the same set of libraries and frameworks are analyzed to recommend methods calls [28] and even the most suitable method parameters [29]. Developers can be supported by recommendation systems, like Hipikat [30], in selecting the correct artifacts of a software project according to the assigned maintenance task. Recommendation systems are also used for architecture comprehension [31], software migration [22], and performance optimizations [32]. Especially in the cloud context, performance and pricing are important aspects to be considered when selecting specific cloud providers for a software system. Therefore, approaches, like CloudAdvisor [33] and CloudMIG [22], support multi-objective optimization. Supplemental, ExplorViz [34], [35] provides a 3D live visualization of the system architecture based on monitoring data to support the human operator to perform changes when adaptations exceed automated planing routines. KAMP [36] supports human operators by semi-automatically analyzing the change propagation in architecture models caused by a given change request, such as replace a database [37] or split an interface [38] for solving a performance bottleneck.

To summarize, while we could not find combinations of automatic and operator-in-the-loop adaptation, SCADA and recommendation systems provide foundations to involve humans in the control loop. Furthermore, there are approaches for architecture visualization and architecture-based change propagation analysis that support operators in evolving the system.

**IV. ADDRESSING THE CHALLENGES**

Addressing aforementioned challenges requires building out a platform for supporting automatic and operator-in-the-loop adaptation processes as depicted in Fig. 1. The cloud application is instrumented with monitoring probes [8]. The Monitor phase exploits information from probes to keep the descriptive architectural run-time models causally connected.
with the underlying cloud application and infrastructure [5]. Architectural run-time models are used during the Analyze phase to determine any problems or deviations and thus triggering the need for an adaptation. Problems may be potential privacy violations or performance bottlenecks. Further, root causes that led to the problem are identified.

Second, constraint-based generation of adaptation alternatives during run-time by including planning activities. First, the introduction of information on external events, and second, the selection of adaptation plans from the design-space exploration. External events, such as sales campaigns, are determined by external parties and cannot be modeled into the architecture model at design-time. Therefore, the operator must be able to introduce new cloud models, data handling policies, and workload models to compose and inject knowledge at run-time. As workload models are complex, operators want to base them on observed behavior. Therefore, the approach will provide the ability to extract and modify intensities and behavior patterns.

The design-space exploration might return multiple adaptation candidates all representing a different element of a Pareto optimum. Human operators must assess and select adaptation plans which is supported by views and analyses. Therefore, we investigate the creation of views on the system that provide operators with up to date information about the actual cloud application derived from the architectural run-time models of iObserve [5], the different adaptation candidates, and the associated adaptation tasks.

V. CONCLUSION

We proposed extensions to the iObserve monitoring and analysis approach for combining automatic and operator-in-the-loop adaptation founded on architectural run-time models, workload characterization, privacy-driven planning, and design space exploration. Extensions are necessary to address challenges when including planning and execution activities.

In the future we will refine the proposed approach and extent the existing iObserve tooling for monitoring and analysis of cloud-based software systems by planning and execution capabilities. We will elaborate and evaluate the approach utilizing the community case studies CoCoME [39] and SPECjEnterprise1. We plan to perform simple user experiments and more complex controlled experiments. In the latter humans will have to execute specific prepared scenarios shaped around a set of external events which partly will occur without preparation for the operator. As base line, we will also perform an automatic selection and application of adaptation candidates.

1SPECjEnterprise is trademark of the Standard Performance Evaluation Corporation (SPEC), http://www.spec.org/JEnterprise2010/