Chapter 20
Self-awareness of Cloud Applications

Alex Iosup, Xiaoyun Zhu, Arif Merchant, Eva Kalyvianaki, Martina Maggio, Simon Spinner, Tarek Abdelzaher, Ole Mengshoel and Sara Bouchenak

Abstract Cloud applications today deliver an increasingly larger portion of the information and communications technology (ICT) services. To address the scale, growth, and reliability of cloud applications, self-aware management and scheduling are becoming commonplace. How are they used in practice? In this chapter, we propose a conceptual framework for analyzing the state-of-the-art self-awareness approaches used in the context of cloud applications. We map important applications corresponding to the popular and emerging application domains to this conceptual framework.
framework and compare the practical characteristics, benefits, and drawbacks of self-awareness approaches. Last, we propose a road map for addressing the open challenges in self-aware cloud and datacenter applications.

### 20.1 Introduction

Cloud computing is the information and communications technology (ICT) paradigm under which services are provisioned by their users on demand, with payment covering only what is actually used. Cloud users can today lease infrastructure, platform, software, and others “as a service,” from commercial clouds such as Amazon, Azure, Google, and SAP. Governments and entire industries are building large-scale datacenters that are and will increasingly host cloud computing applications. At the same time, data become an integral part of cloud computing: By 2017, over three quarters of our personal and business data will reside in datacenters, according to a recent IDC report [32]. Cloud applications, often consumed by users as services, already represent over 10% of the entire ICT market in Europe [15]. Netflix, whose users consume a large fraction of the global Internet traffic, relies on ICT services from Amazon Web Services (AWS).\(^1\) The market, growing in size, diversity of applications, and sophistication, already exceeds hundreds of millions of users and, as a consequence, $100 billion worldwide [10]; the cloud market will likely contribute over 100 billion euro to the European GDP in 2020 [15]. At this scale and with this importance, human management of datacenter resources is prohibitively expensive and, often, too error-prone. Thus, the use of self-awareness techniques to manage cloud applications is increasingly more present. In this chapter, we analyze the use of self-awareness in cloud computing and its applications.

Cloud applications raise a complex management challenge, derived from the goals of three main stakeholders: application users, application operators, and cloud operators. Each of these stakeholders has different requirements, which are often conflicting. For example, application users could demand that an interactive application is always responsive, even under bursty arrivals of user-issued requests. To meet this demand, application operators could require that enough capacity is always provided by cloud operators, yet only want to pay for what is actually consumed. Tension arises between performance and other requirements, including the cost of operations. As a consequence, the management challenge is to optimize non-trivial efficiency metrics and to meet complex service-level agreements (SLAs), to an extent that already exceeds the capabilities of human management.

We investigate in this chapter the current state of self-awareness in cloud computing and in particular datacenter-based cloud applications. Our goals are to introduce practical cases of self-awareness in such applications; to present a conceptual framework for analyzing state-of-the-art self-awareness approaches used in practice; to map the already important and emerging application domains to the conceptual

---

\(^1\)Details: [https://aws.amazon.com/solutions/case-studies/netflix/](https://aws.amazon.com/solutions/case-studies/netflix/).