Chapter 14
Mobile Cloud Resource Management

Konstantinos Katzis
European University Cyprus, Cyprus

ABSTRACT

Providing mobile cloud services requires seamless integration between various platforms to offer mobile users optimum performance. To achieve this, many fundamental problems such as bandwidth availability and reliability, resource scarceness, and finite energy must be addressed before rolling out such services. This chapter aims to explore technological challenges for mobile cloud computing in the area of resource management focusing on both parts of the infrastructure: mobile devices and cloud networks. Starting with introducing mobile cloud computing, it then stresses the importance of resource management in the operation of mobile cloud services presenting various types of resources available for cloud computing. Furthermore, it examines the various types of resource management techniques available for mobile clouds. Finally, future directions in the field of resource management for mobile cloud computing environment are presented.

INTRODUCTION

Mobile devices such as tablets, phablets, smartphones etc. are increasingly becoming an essential part of our daily life as the most effective and convenient tool regarding communication, work, general knowledge access, maps, diary, health monitor, etc. The explosion in numbers of such devices increased mobile broadband coverage and as a result the need for always-on collaborative services. Microsoft CEO Satya Nadella stated in his keynote speech at Worldwide Partner Conference 2014 in Washington DC on the 16th of July 2014 that there is a "tremendous opportunity" with mobile and cloud solutions and that Microsoft is focused on defining the "next-generation of productivity" (Nadella, 2014). He also stated that there are going to be more than 3 billion people with connected devices on top of 200 billion sensors all of which will generate tons and tons of zettabytes of data and they are going to consume and reason over the large data. It is true that there has been significant advances in cloud computing and mobile technologies both changing our lives, the way we do business and the way we communicate and this is primarily because of the rapid scal-

DOI: 10.4018/978-1-4666-8751-6.ch014

ability of cloud services, the ubiquitous network access, on-demand self-service etc. Mobile users experience on a daily basis various mobile cloud services, which either run on their devices and / or remote servers with both data storage and data processing happening outside their mobile device. However, this experience is highly dependent on how cloud-resources are managed.

A cloud-computing infrastructure as described by Pallis (2010), is a complex system that is constantly evolving, consisted of numerous devices (infrastructure, platforms and software) with a great deal of requirements, aiming towards the evolution and convergence of several independent computing trends such as internet delivery, distributed computing, storage, content outsourcing, security etc. Cloud computing is based on a large number of shared resources that are subject to spatial and temporal based varying resource requests. The nature of resource requests varies thus making resource management a complex task. Resource management can play a decisive role in optimizing the performance of such a complex system by efficiently and fairly managing performance, functionality and cost. Inefficient resource management can result in negatively affecting the performance and cost while it may also impair system functionality. Due to its complexity and inhomogeneity in the type of services it offers, cloud resource management requires complex policies and decisions to fulfil a multi-objective optimization. This is extremely challenging because of the vast scale of the system, which makes it impossible to have accurate global state information that are not subjected to unpredictable interactions. According to Marinescu (2013), the main strategies for cloud resource management are associated with the three types of cloud delivery models. These are: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). In all cases, the cloud services providers are faced with large, fluctuating loads that sometimes are unable to serve.

Cloud Computing envisages that future computer applications and services are constructed, delivered and managed in a much different way than what we are used to so far. Cloud Computing aims to move computing and data away from the common desktop and portable computers to allow the use of smaller, more portable and less power demanding devices. According to Dikaiakos et al. (2009), there are two different architectural models being identified. The first one is designed to scale out by providing additional computing instances for supplying SaaS and PaaS services and the second is designed to provide data and compute-intensive applications via scaling capacity. Setting the physical layout of a cloud infrastructure can vary significantly since there isn't a specific recipe to be followed. More specifically, a cloud can be restricted to single organization or a group as a private cloud or it can be available to the general public as a public cloud or something in-between as a hybrid cloud. In essence, any type of cloud is comprised of three basic elements. The processing, network and storage element each of which defines the level of quality of service required by the customers. This is presented by Dikaiakos et al. (2009) and illustrated in Figure 1 as a cloud architecture consisted of 3 layers. Starting from bottom up first comes the infrastructure, which involves storage and computing capabilities. Second comes the platform layer that provides the means for developing, testing, deploying, hosting and maintaining applications in an integrated environment that is the cloud. The application layer is the last one, which aims to offer a complete application as a service.

Performance in cloud computing is subject to how well the hardware and software are integrated and managed. Current trends show that computer architecture is shifting towards multiple core processors to address the increasing levels of parallelism whereas software is constantly redesigned by academia and industry for parallel and data-intensive computing. A generalized version

Related Content