Journal of Network and Computer Applications 70 (2016) 102–113

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Cloud based Video-on-Demand service model ensuring quality of service and scalability

Carlos Barba-Jimenez [a], Raul Ramirez-Velarde [a,*], Andrei Tchernykh [b], Ramón Rodríguez-Dagnino [a], Juan Nolazco-Flores [a], Raul Perez-Cazares [a]

[a] Tecnologico de Monterrey, Campus Monterrey, Ave. Eugenio Garza Sada 2501, Monterrey, N.L. 64849, Mexico
[b] CICESE Research Center, Carretera Ensenada-Tijuana 3918, Ensenada, B.C. 22860, Mexico

ARTICLE INFO

Article history:
Received 7 September 2015
Received in revised form
29 February 2016
Accepted 10 May 2016
Available online 13 May 2016

Keywords:
Video-on-Demand
Cloud computing
Elasticity
Heavy-tails
PCA
Self-similarity

ABSTRACT

Increasing availability and popularity of cloud Storage as a Service (STaaS) offers alternatives to traditional on-line video entertainment models, which rely on expensive Content Delivery Networks (CDNs). In this paper, we present an elastic analytic solution model to ensure Quality of Service (QoS) when providing Video-on-Demand (VoD) using several third party elastic cloud storage services. First, we individually gather cloud storage start-up delays, and characterize them to show that they are heavy-tailed. Then, we perform a meta-characterization of these delays using Principal Component Analysis (PCA) to create a characteristic cloud delay trace. By using different estimation techniques of the Hurst Parameter, we demonstrate that this new trace (also heavy-tailed) exhibits self-similarity, a property not sufficiently studied in cloud storage environments. Finally, we pursue stochastic modeling using different heavy-tailed probability distributions to derive prediction models and elasticity parameters from the cloud VoD system. We obtain a stochastic self-similar model and compare it with trace based simulation results by testing different heavy-tailed probability distributions, meta-cloud elasticity values and Hurst parameters. Since our approach optimizes QoS, we guarantee a specific video start-up delay for a number of arriving clients. This is a strong commitment for a VoD service, because traditional cloud approaches often focus on a best-effort paradigm optimizing performance, cost, and bandwidth, among other parameters.

1. Introduction

Nowadays, the cloud paradigm has become increasingly popular, offering new services and products every year. The concept has reached areas beyond traditional IT environments, research, software development, and even entire business models. Providers such as Amazon offer elastic computing and cloud Storage as a Service (STaaS) commercially. These services have increased interoperability, usability and reduced cost of application hosting, content storage and delivery (Buyya et al., 2010). These conditions open the door for creating new services that can satisfy existing and future user demands. This is especially important if one considers the trends of IP traffic reported in Cisco (2013), where it is shown that Video-on-Demand (VoD) traffic will be tripled in 2017 with rising mobile content consumption (Passarella, 2012).

VoD and online video content services are generally supported by a centralized delivery architecture based on private or rented servers with fixed costs and little flexibility (Buyya et al., 2009). Such a model poses a challenge, since predicting the user demand erroneously could cause performance bottlenecks with an under-estimation, and high costs with an over-estimation. The model has to be adapted to include a Content Delivery Network (CDN), usually operated by a third party in multiple external sites (Buyya et al., 2009). CDNs have a central entity that can enforce Quality of Service (QoS) levels, but this comes at a non-negligible financial cost (Passarella, 2012).

There are Peer-to-Peer (P2P) alternatives, but rarely provide guaranteed services (Passarella, 2012). They are used primarily in live streaming, where they help with flash crowds that need to access certain content at the same time (Mansy and Ammar, 2011). This is not always the case in a VoD service. Now, the CDN model is still the most prevalent, with services like Akamai. The recent P2P propositions in Thomas et al. (2015) have included a hybrid cloud component to facilitate video streaming by taking advantage of characteristics from both technologies, but only focus on a best-effort approach to QoS.

There are third party services (e.g. YouTube, Vimeo, etc.) for delivering video content. However, even with large infrastructures behind, they only offer best-effort QoS. It makes them not suitable

* Corresponding author.
E-mail address: rramirez@itesm.mx (R. Ramirez-Velarde).

http://dx.doi.org/10.1016/j.jnca.2016.05.007
1084-8045/

for all businesses and entertainment conditions.

STaaS can be used as an alternative, with some commercially available services that use characteristics and advantages of the cloud, in a similar manner as proposed under the MetaCDN project (Broberg et al., 2009). The authors describe characteristics of the model, but left out details of the algorithms. The described statistical analysis is also limited for accurate predictive models. Our work has a similar motivation.

We extend the preliminary work presented in Barba-Jimenez et al. (2014) by widening the scope of our study and the problem definition. We include new parameters and characteristics to the proposed end-to-end VoD service. Furthermore, in this paper, we consider the derivation of the different stochastic models for predicting user capacity under certain QoS level and start-up delay taking into account the self-similarity property of the Characteristic Cloud Delay Trace (CCDT), and elasticity characteristics of the cloud. Additionally, we compare and evaluate the model and uncertainties of its parameters against simulation results. These aspects were not fully elucidated neither in Broberg et al. (2009) nor in Barba-Jimenez et al. (2014).

In this paper, we use third party cloud services to create a meta-VoD service (similar to MetaCDN) using a gateway as suggested in Islam and Grégoire (2012). This gateway creates a CDN like functionality sitting in a layer above the clouds. It takes into account some of the main challenges of VoD content delivery, namely, response time, and start-up delay.

As a metric for the meta-VoD service quality, we use start-up delay. We take into account the abandonment rate described in Krishnan and Sitaraman (2012), where the authors found that VoD clients start leaving the service after a start-up delay of 2000 ms (milliseconds), losing 5.8% of users for each additional second. This delay time T_D includes all network times, server times, and additional overheads.

We propose a methodology to model and analyze storage cloud delays, considering their statistical characteristics, and elasticity. Then we develop a meta-VoD elastic model that estimates the number of users that can be served for a given QoS and start-up delay time. This model also uses self-similarity and heavy-tail properties, following Ramirez-Velarde et al. (2013) and Ramirez-Velarde and Rodríguez-Dagnino (2010).

We provide background of our work and introduce an elasticity concept in Section 2. We discuss related work in Section 3. Then, in Section 4, we present the basic solution model. Section 5 includes a detailed statistical analysis regarding real cloud data delay traces. It includes the determination of heavy-tails. We characterize real cloud data using statistical analysis, determine the heaviness of each individual cloud delay time tails and then reduce dimensionality of these data sets using Principal Component Analysis (PCA). We provide a meta-characterization of the individual clouds with the CCDT. The objective is to simplify the model for the delay time T_D, which enables us to make predictions of the probability of successful service under a certain threshold of abandonment rate. Section 6 introduces the concept of self-similarity and different estimations using the CCDT. Additionally, the self-similarity and elasticity are included to derive delay stochastic models using sub-exponential probability distributions. Section 7 validates the models by experimentation. Sections 8 and 9 describe the results and conclusions.

2. Background

2.1. Cloud computing

The cloud is defined as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort and service provider interaction. Cloud computing has three service models: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) (Mell and Grance, 2009).

Cloud providers offer service level agreements (SLAs), which guarantee a level of QoS. Resource usage is monitored, controlled and reported, providing transparency for both the provider and consumer (Espadas et al., 2013). This makes the idea of using the existing third party cloud services for content distribution very attractive. Paying only for storage and computing, instead of having a potentially expensive contract with one CDN or an under-provisioned private server.

2.2. Cloud elasticity

One of the main characteristics of cloud computing is the pay-per-use model. In order to provide metered services and resources under SLAs, the cloud providers must be able to match the resource demand with the resource offer as close as possible. Elasticity can be defined as the degree to which a system is able to adapt to workload changes by provisioning and de-provisioning resources in an autonomic manner, such that at each point in time the available resources match the current demand as closely as possible (Herbst et al., 2013).

In real scenarios, clouds are not perfectly elastic (Brebner, 2012). The infrastructure cannot respond instantly to sudden, significant increases in demand. There is a delay between the time when resources are requested, and when the application starts running.

Elasticity has been studied in several works, including Almeida et al. (2013), Costa et al. (2013), Herbst et al. (2013) and Kaur and Chana (2014), for frameworks considering costs, QoS, under/over provisioning and task execution. Ideally, there would be an external way of measuring or polling a measure of elasticity at any given time. However, obtaining these values from outside the cloud black-box is not easy.

We denote this elasticity metric as ξ, where $0 < \xi \leq 1$. In this definition, 1 describes a 100% elastic cloud system, where the resources always match the demand in every instant in time. The proposed ξ metric is similar to the precision of scaling up described in Herbst et al. (2013).

3. Related works

We address the video CDN based on the cloud. In Broberg et al. (2009), the authors present the price comparison of delivering content through a CDN versus different cloud providers. The traditional CDN model is the most expensive in TB data/month, while cloud and cloud CDN options come as the cheaper alternatives. The authors presented interesting points related to QoS provisions. However their proposal is vague for a proper mathematical setting.

In a recent work, the use of cloud CDN-like functionality has also been reported (Guan and Choi, 2014). However, it is aimed at minimizing bandwidth and cost in the content placement problem (from a provider point of view). In contrast, we consider the latency and QoS as user centric criteria.

Additionally, on the topic of CDN modeling, in Buyya (2009), the authors explore resource discovery and request redirection in a multi provider content delivery network environment. They show that CDNs evolve as a solution for Internet service degradations and bottlenecks due to large user demands to certain web services. They address some of the internal problems that CDN providers