Exploring Inter-Cloud Load Balancing by Utilizing Historical Service Submission Records

Stelios Sotiriadis, University of Derby, UK
Nik Bessis, University of Derby, UK
Nick Antonopoulos, University of Derby, UK

ABSTRACT

Cloud computing offers significant advantages to Internet users by deploying hosted services via bespoke service-provisioning environments. In advance, the emergence of Inter-Cloud increases the competences and opportunities of clients for a wider resource provision selection. This extends current capabilities by decoupling users from cloud providers while at the same time cloud providers offer an augmented service delivery mean. In practice, cloud users make use of their brokering component for selecting the best available resource, in terms of computational power and software licensing of a datacenter based on service level agreements for service execution. However, from the cloud perspective, the overall choice for balancing the different workloads within the Inter-Cloud is a complex decision. This article explores the performance of an Inter-Cloud to measure the utilization levels among their sub-clouds for various job submissions. With this in mind, the solution is modeled for achieving load balancing based on historical records from past service execution experiences. The record files are composed in the form of log files that keep related information about the size of the Inter-Cloud, basic specifications, and job submission parameters. Finally, the solution is integrated in a simulated setting for exploring the performance of the approach for various heavy workload submissions.

Keywords: Cloud Computing, Cloud Load Balancing, Cloud Scheduling, Historical Records on Job Delegations, Inter-Cloud

INTRODUCTION

A cloud-computing environment includes the delivery of hosted services that are located in a remote location via the public Internet to everyday users. Although various opinions include a narrow view of clouds – mainly as an enterprise server-based datacenter – in this article we have taken an inclusive perspective of clouds as to encompass various kinds of resources. At a first glance, cloud computing share similar fundamental elements with other large scale and/or distributed computing paradigms e.g., clusters and grids. In addition, a broad definition of cloud computing includes a service-oriented virtualization environment. This generic vision – from the perspective of the service – change the focus on how to or-
orchestrate the cloud service distribution, rather than aim to the management and deployment of the underlying infrastructure.

In such environments, massive computing capacity resides at a remote space and could be delivered in the form of software and/or hardware (Carolan & Gaede, 2011). These offered services are identical to job submissions that have been encapsulated in application execution requests that have been posed by the end-users. Although cloud computing is still in its infancy due to the facility-orientation (physical space), it needs to be evolved to a more distributed infrastructure with a broader propagation of services. This could be achieved by utilizing available resources (clusters, high performance computing and grids) relying at the lower level (infrastructure). By transforming the cloud infrastructure to go beyond its premises we could facilitate a wider set of deployed services and applications.

The study aims to the InterCloud load balancing which represents an interconnected global cloud of clouds (Sotiriadis et al., 2012a). The generic idea as presented by Bessis et al. (in press-b) and Buyya et al. (2010) is to decouple resource consumers from providers and allowing providers to offer resources on demand and on an ad hoc basis. For achieving this model, a new structure should be established to contain the required conceptions of resource decoupling. These are protocols to control trust standards, discovering, systems for naming, scheduling of services, portability and workload exchange. We focus majorly on the service management concept, and eventually the load balancing mechanism during the service submissions. The target is to effectively achieve an enhanced quality of service by methodologically assign services in the form of job tasks (processes) to resources. These services (jobs) encapsulate various capabilities of the cloud environment (e.g., provisioning of software and/or hardware). Thus, the challenge is to identify the rationality behind the decisions of the cloud provider to manage the Inter-Cloud service execution for efficient load balancing.

Specifically, the users submit their requests in a broker with the latter communicating and monitoring the whole service exchanging procedure. This component is responsible for autonomous decisions by selecting a datacenter for forwarding the request. Then, each request is sandboxed in a virtual machine (VM) that satisfies these requirements. Various criteria are implemented in this level that includes the user-defined quality of service levels e.g., pricing, homogeneity in terms of hardware and software, and generic specification of the cloud VM. These are enclosed in service level agreements (SLAs) that formally define the level of agreed terms between provider and client. Usually, this is related with the required computational power (performance) and time constraints.

To this extend, we have presented a meta-brokering solution in Sotiriadis et al. (2012b) as a novel component that is placed on the top of each broker. The aim was to achieve the decentralization of the setting in which meta-brokers collaborate with each other for SLAs trading. By using cloud meta-brokers an InterCloud is formed into an autonomously management setting of interconnected sub-clouds. Current efforts in this direction organize (meta-) centralized topologies of brokers, so various drawbacks derived from this narrow view. Herein, the work is inspired from the meta-computing concept and based upon the model of a decentralized meta-broker. Specifically, we measure the utilization level of the Inter-Cloud and we save results in log files named as historical records. Then, we explore the performance of the Inter-Cloud for various sub-cloud numbers and job variations.

With this in mind, the following section presents the motivation of our work, which is related, with the area of large-scale systems. The rest of the paper is organized as follows; we present the InterCloud load balancing solution that is followed by the presentation of the experimental analysis and results of the Inter-Cloud utilization levels and the load-balancing mechanism. At last we conclude our work by
Service Level Agreements for Real-Time Service-Oriented Infrastructures
www.igi-global.com/chapter/service-level-agreements-real-time/64570?camid=4v1a

Probabilistic Indices of Quality of Approximation
www.igi-global.com/chapter/probabilistic-indices-quality-approximation/28472?camid=4v1a