Future Generation Computer Systems 26 (2010) 947–970

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Toward dynamic and attribute based publication, discovery and selection for cloud computing

Andrzej Goscinski *, Michael Brock

School of Information Technology, Deakin University, Pigdons Road, Waurn Ponds, 3217, Australia

ARTICLE INFO

Article history:
Received 3 December 2009
Received in revised form
9 March 2010
Accepted 27 March 2010
Available online 14 April 2010

Keywords:
Attribute based publication
Service discovery and selection
Ease of use
Clouds
Cloud computing

ABSTRACT

Cloud computing is an emerging paradigm where computing resources are offered over the Internet as scalable, on-demand (Web) services. While cloud vendors have concentrated their efforts on the improvement of performance, resource consumption and scalability, other cloud characteristics have been neglected. On the one hand cloud service providers face difficult problems of publishing services that expose resources, and on the other hand cloud clients do not have the means for discovery and automatic services' selection, and easy use of services.

In response, proposed in this article is the application of the Resources Via Web Services framework (RVWS) to offer higher level abstraction of clouds in the form of a new technology. Our new technology makes possible the provision of service (and resource) publication, discovery and selection based on dynamic attributes which express the current state and characteristics of cloud services and resources.

A proof of concept implementation that allowed the easy publication, discovery, selection and use of an existing cluster (one of the most frequently used cloud resource) via a simple interface using Web pages backed by extensive sets of tests has demonstrated that the design is sound and the proposed technology is feasible. The proposed solution is significant: instead of spending time and effort locating, evaluating and learning about clusters, clients are able to easily discover, select and use the required resources. Furthermore, service providers (which can be entities external to clouds themselves) can easily publish (and keep current) information about their services (and the resources behind them).

1. Introduction

The future of computing lies in cloud computing, whose major goal is reducing the IT services' costs while increasing processing throughput and decreasing processing time, increasing reliability, availability and flexibility [1]. Cloud computing is a new paradigm where computing resources (from data storage to complete configurations of distributed systems) are made available (offered) over the Internet as scalable, on-demand (Web) services. Diagram 1 shows a general view of clouds (and their resources) in relation to a client.

In cloud computing, the resources hosted within clouds can be anything: they could be database services, virtual servers (virtual machines), complete service workflows or complex configurations of distributed computing systems such as clusters. Regardless of their nature, all resources are provided via services to clients (users or software processes) by computers rented from the cloud (such as those offered by e.g., Amazon, Google, Microsoft), rather than by private systems. The services are provided on demand and clients only pay for the quantity of resources (data storage, computation, etc.) they use.

In addition to services and resources, cloud computing has providers of two forms: *service providers* and *cloud providers*. A cloud provider is the entity that offers and maintains a cloud and may offer internally developed services on the cloud. A service provider is an entity that creates and maintains services that are published in and ran on clouds. For example, a service provider may not have the capita to host its own services hence rents space on a cloud to reduce costs.

An analysis of products and services released by Microsoft [2], Amazon [3], Google [4] and Salesforce [5–7], the best known cloud service providers, shows that clouds fall into a number of categories: applications (*Software as a Service—SaaS*), platform (*Platform as a Service—PaaS*) and hardware (*Infrastructure as a Service—IaaS*).

In the SaaS category, there is delivery of use-specific services over the Internet (such as CRM software and email). The benefit of SaaS clouds is that clients only focus on the use of the software and do not have to worry about the cost and effort to keep software licenses current nor the handling of software updates. However, SaaS clouds are not without their risks. The decision on whether or not to deploy software updates are finalized by the providers themselves. Thus, if an update to a software services makes a client

* Corresponding author.
E-mail addresses: ang@deakin.edu.au (A. Goscinski), brock@deakin.edu.au (M. Brock).

0167-739X/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2010.03.009

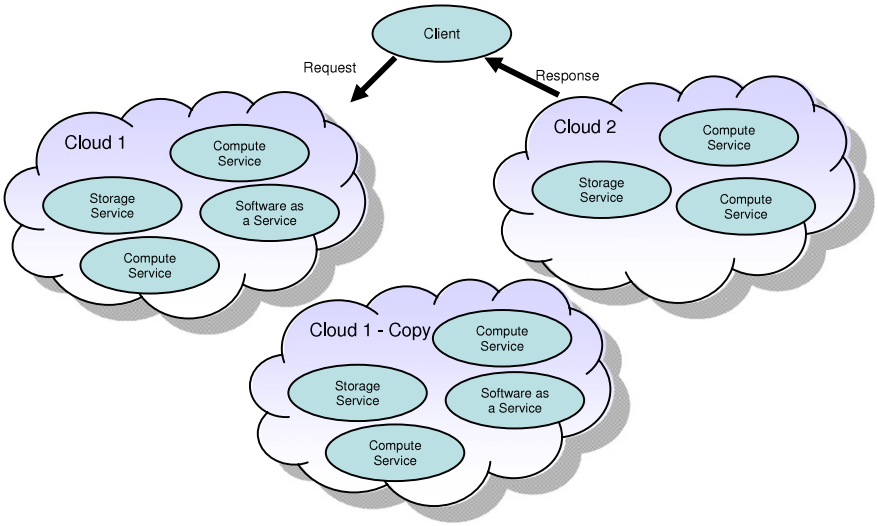

Diagram 1. A single client vs. services provided by computing clouds.

incompatible, the client has to either adapt their software or find another software service or even find another SaaS cloud.

The PaaS category represents clouds that access to a range of compute, database and storage functions within a specified framework provided over the Internet. The benefit of PaaS clouds is clients are able to create their own required services and do not have to worry about provisioning and maintaining the hardware and software needed to run the services. On comparison, PaaS is like SaaS except clients are able to create software as well as use it.

Finally, the IaaS category allows for the provisioning of hardware resources so cloud clients can create various configurations of computer systems from servers to complete clusters. Clients to IaaS are able to host their own services and even complete software systems without having to worry about hardware costs. On comparison to PaaS and SaaS, clients are able to create and use software as well as create and use an underlying software infrastructure to make the software possible.

While there is flexibility in what resources are provided, the underlying clouds are proprietary and specialized for internal uses. For example, while Amazon EC2 is the most commonly known IaaS cloud, how EC2 discovers and selects required resources from among their virtual servers (memory, CPU cycles, etc.) is not publicly known nor if such a function exists at all. There is also another issue: businesses experience problems when they wish to make services of their private clouds available.

While vendors have concentrated their effort on the improvement of performance, resource consumption and scalability, other cloud characteristics have been neglected. Clients face difficult problems of resource discovery and automatic services selection; dynamic sharing toward efficient management of resources; QoS and reputation of providers and clients; fault tolerance; cloud security and ease of use have been neglected. For example, clients to IaaS clouds are heavily involved in configuration of virtual servers and the execution of their applications in the same manner as programmers did years ago when they used a Unix system [8,9].

The ineffectiveness of current cloud discovery is made worse by the fact that resources in clouds are not confined to the storage of data or the creation of virtual servers. One of the most frequently used resources of clouds is a cluster. Even if clouds had an effective means of discovery, publishing services that expose clusters is not easy. Currently, resources are exposed using stateless Web services hence the state is not publishable. Clusters are complex and change their state frequently over time; hence exposing clusters via a service for publication is itself a challenge. With the emergence of cloud computing, high level simplification of cluster abstraction is vital. This is becoming more urgent these days when clouds are being adapted to support High Performance Computing (HPC) [10–15]. Thus, there is yet to be a discovery service that easily and effectively finds clusters on the Internet cloud for research users.

Even when a cluster has been found, using one is not easy. As most clusters are built using Unix-like software, there is a significant lack of user ease of use. A quick look at leading cluster middleware [16–20] has found that while there is effective allocation of processes to cluster nodes, little attention has been made to the ease of use of clusters and their middleware facilities. While attempts have been made to try and find a Web based solution for access clusters, only a single solution from 2003 could be found [21,22].

The goal of this article is to present a solution that makes discovery, selection and use of cloud services and resources (among them clusters) easy. Proposed in this article is the use of the innovate, dynamic attribute and Web service based Resources Via Web Services (RVWS) framework [23,24] in building an easy to use, Web based technology called the Cluster as a Service (CaaS) Technology. Through the use of our CaaS Technology, publishing, discovering, selecting and using a required cluster is easy and effortless.

The contribution of this article is in two parts: (1) an innovative, general framework for service (and resource) publication, discovery, selection and use via dynamic attributes that expose current state and characteristics via Web services and its enhancement for one special cloud's resource, a cluster, in the form of a Cluster as a Service (CaaS) Technology; and (2) a proof of concept implementation and testing that demonstrate the feasibility and correctness of our framework and technology.

The significance of the outcome of our research presented here is a means for the creation of higher layer abstraction of clouds, shown here as a for cluster oriented technology that offers easy publishing, discovery, selection and use of cloud's clusters through a Web based interface supporting both clients wishing to use cloud services and resources and service providers who wish to publish their own services and resources.

The rest of this article is structured as follows. Section 2 examines related work in the areas of clouds, grids and clusters: the focus of which is the presentation of the state of the art in service/resource publication, discovery, selection and use. Section 3 gives a brief explanation of RVWS, mainly its exposure of dynamic attributes. Section 4 presents the logical design of the CaaS Technology, including an intuitive and easy to use interface