Chapter 13

An Infrastructure-as-a-Service Cloud: On-Demand Resource Provisioning

Weijia Song
Peking University, China

Zhen Xiao
Peking University, China

ABSTRACT

Cloud computing allows business customers to elastically scale up and down their resource usage based on needs. This feature eliminates the dilemma of planning IT infrastructures for Cloud users, where under-provisioning compromises service quality while over-provisioning wastes investment as well as electricity. It offers virtually infinite resource. It also made the desirable “pay as you go” accounting model possible. The above touted gains in the Cloud model come from on-demand resource provisioning technology. In this chapter, the authors elaborate on such technologies incorporated in a real IaaS system to exemplify how Cloud elasticity is implemented. It involves the resource provisioning technologies in hypervisor, Virtual Machine (VM) migration scheduler and VM replication. The authors also investigate the load prediction algorithm for its significant impacts on resource allocation.

1. INTRODUCTION

Cloud elasticity refers to the ability of Cloud infrastructure to dynamically make resource provision for Internet applications and services, according to their real time requirements. That feature of Cloud Computing has several appealing implications. It eliminates the dilemma of planning IT infrastructures for Cloud users, where under-provisioning compromises service quality while over-provisioning wastes investment as well as electricity. It offers virtually infinite resource. It also made the desirable “pay as you go” accounting model possible.

Planning new IT infrastructure for growing demands of Internet applications is complicated. It calls for successful prediction on how appli-
An Infrastructure-as-a-Service Cloud

cation loadings would change in the future and is particularly hard for Start-Ups since market response is not clear in advance. When it knocks against flash-crowd, self-maintained servers may fail to satisfy the need of surging requests. On the contrary, over provisioning caused by optimistic prediction leaves the server under-utilized, and consequently causes waste in energy and excess investment in fixed asset. In Cloud environment, however, application maintainers need not worry about such problems since the Cloud resource allocation automatically scales up and down on changing load, and the users are billed accordingly.

Sometimes, particularly in data mining applications, a user may require a large number of servers for a short period. It is hard to satisfy such requirement if it were not for Cloud computing. A successful example is, “The Washington Post uses Amazon EC2 to turn Hillary Clinton’s White House schedule—17,481 non-searchable PDF pages—into a searchable database within 24 hours.” (“AWS Case Study: Washington Post,” n.d.). In colleges, researchers may have similar requirements when processing huge amount of experiment data.

There are different approaches to Cloud elasticity depending on how the Cloud infrastructure is constructed and what types of applications running over it. In the next sections, we are going to introduce some popular technologies adopted nowadays Cloud Services. Then we start from basic components of a Cloud infrastructure to explain our own work that handling Cloud elasticity in a real IaaS service. In the end, we will point out, in our perspective, the trend of Cloud elastic technologies.

2. BACKGROUND

Traditionally, Cloud services are categorized into Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). IaaS users are responsible for application development, deployment and management. PaaS take the burden of application management by providing development tools and deployment platform. SaaS model is actually “old wine in new bottles” for conventional Internet applications.

In an IaaS system, virtual machines are generally overcommitted to physical servers to maximize profit from hardware investment and cut down power budget. Cloud elasticity in that environment addresses the challenge of resource provisioning for dynamic load of virtual machines. For example, if physical server cannot satisfy the resource requirements of its virtual machines, some of them are going to be migrated to other servers so that application performance is assured.

It is hard for application developers to predict the user load. In PaaS systems, user applications are managed by Cloud infrastructure to relive developers of the difficult of deployment, so that they can concentrate on application function. Generally, the applications deployed in PaaS are developed by designated program language, development tool and libraries and encapsulated in managed execution engines. An execution engine is a sandbox allocated with a share of CPU resource. Execution engines have uniform management interface for life cycle control and performance monitoring. Elasticity mechanism dynamically adjusts the number of execution engines belonging to an application to suit its load.

The situation in SaaS is similar to that in PaaS. A SaaS service could be built upon a PaaS service to indirectly utilize its elasticity mechanism. Some large SaaS services choose to implement dedicated elasticity mechanism for application specific optimizations (Chen et al., 2008) (Chase et al., 2001). Here we just talk about stateless computing resource. The discussion of data storage technologies like Google File System (Ghemawat, Gobioff, and Leung, 2003) and Big Table (Chang et al., 2008) belongs to another dedicated field out of scope of this chapter.
Related Content

Virtual Tutoring: The Case of TutorVista
[www.igi-global.com/chapter/virtual-tutoring-case-tutorvista/44010?camid=4v1a](www.igi-global.com/chapter/virtual-tutoring-case-tutorvista/44010?camid=4v1a)

Service Platform Development: Comparison of Two E-Services Platforms
[www.igi-global.com/article/service-platform-development/53230?camid=4v1a](www.igi-global.com/article/service-platform-development/53230?camid=4v1a)

Data Partitioning for Highly Scalable Cloud Applications
[www.igi-global.com/chapter/data-partitioning-highly-scalable-cloud/65297?camid=4v1a](www.igi-global.com/chapter/data-partitioning-highly-scalable-cloud/65297?camid=4v1a)

Aiding Compliance Governance in Service-Based Business Processes
[www.igi-global.com/chapter/aiding-compliance-governance-service-based/60900?camid=4v1a](www.igi-global.com/chapter/aiding-compliance-governance-service-based/60900?camid=4v1a)