Chapter 1

Literature Review

1.1 Background

This chapter begins by providing an overall understanding of Cloud computing and its related research field. Then, it narrows down to the server consolidation problem in Section **??**.

1.1.1 An Overview of Cloud Computing

Cloud computing allows their users to access Cloud resources from anywhere in the world. Software developers deploy their softwares in the Cloud in a form of service, hence, their customers can use them without installing on their local computers. Cloud computing has made one critical change in software industry, it separates the role of traditional service provider into service provider and infrastructure provider. As Wei [**?**] states, "one provides the computing of services, and the other provides the services of computing". Therefore, this separation add one more layer between service provider and users, as: Cloud provider, Cloud users (service provider), and End users.

Each of these stakeholders has their goal. End users consume the application deployed on Cloud. They require a guarantee quality of the software including functional requirements which is an expected functionality, and non-functional requirements which are addressed as performance requirements such as availability, security, and network latency.

Cloud users deploy their software on Clouds. They want to increase the profit by increasing income and decreasing expense. Increase of income is mainly through two ways, either attract more End users or increase the charge. In order to achieve these two objectives, they can improve the functionality of the software, or improve the non-functionality features by guaranteeing Quality of Service (QoS).

To improve the non-functionality features, service capacity planning is the core process. The capacity planning has two conflicting objectives, on one hand, it must meet End users' QoS requirement by using enough resources. On the other hand, the cost must be minimized. In pre-Cloud era, the capacity planning determines the upfront investment in infrastructure, therefore, capacity, reliability, and scalability are all need to be carefully considered and balanced. In Cloud environment, the burden of capacity planning is largely released by elastic resource management and the pay-as-you-go policy.

Cloud users identify a list of critical QoS parameters called Service Level Agreement (SLA) which specifies the non-functional requirements such as throughput, latency, and availability. These QoS parameters are mapped to resources (e.g. CPU, memory, network bandwidth) which can satisfy these requirements. Violation of SLA will lead to penalty and decreasing in number of users. Therefore, in essence, the key to attract more users is an

effective resource management system which can rapidly react to the fluctuating resource demand.

Beside increase the income, reduce the expense is another way to improve profit. As previous section mentioned, energy consumption is the main source of expense. In energy consumption, server energy consumption is the core that needs to be improved.

1.1.2 Energy-aware Resource Management

1.1.3 An Overview of Evolutionary Computation

In order to understand Cloud computing, firstly we will illustrate the five essential elements of Cloud computing and their advantages.

Cloud computing has five essential elements:

1. On-demand self-service, it means a Cloud user can require computing resources (e.g CPU time, storage, software use) without the interaction with Cloud provider.

2. Broad network access, Computing resources are connected and delivered over the network.

3. Resource pool, a Cloud provide has a "pool" of resources which are normally virtualized servers. In IaaS, it provides predefined sizes of VMs. In PaaS, the resources are 'invisible' to Cloud users who have no knowledge or ability to control.

4. Rapid elasticity, from the perspective of Cloud users, computing resources are assigned and released in real time. In addition, the resources assign to their software is "infinite". Therefore, Cloud users do not need to worried about the scalability of their applications.

5. Measured Service provides an accurate measure of the usage of computing resources. It is fundamental to the pay-as-you-go policy.

What is your purpose to describe the following content? I would like to discuss the differences, advantages of disadvantage of the resource management in different service models. Therefore, after illustrate how they are work. The point is to compare the resource management. And then, lead to a new service model. And the advantage of new service model should be obvious.

Traditional Cloud computing has three service model as illustrated in Figure **??**.

1. Infrastructure as a Service, Cloud provider offers the fundamental computing resources, often in the form of various sizes of VMs. Apart from the virtualized hardware and operating systems, Cloud users treat the remote servers as local and deploy their applications. In terms of resource management, Cloud users have the responsibility to estimate the quantity of resources, while Cloud providers have no knowledge and control inside VMs, resource management is based on VM.

2. Platform as a Service, Cloud providers establish the software development platform to enable the in-progress software to be developed in the platform. The main difference between PaaS and SaaS is that, PaaS supports the full life cycle of software development, whereas SaaS only host completed applications deployment. In terms of resource management, Cloud providers have the full control of resource allocation, auto-scaling and consolidation. Therefore, Cloud users can focus on software development.

3. Software as a Service (SaaS). Cloud users deploy their applications in Cloud which can be accessed by End Users. SaaS describes the relationship between Cloud provider and End users with the connection of applications.

Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service.

1.1.4 Resource Management

Scope of Cloud computing resource management.

1. Actors
2. Management Objectives
3. Resource Types
4. Enabling Technologies

Energy-aware Resource Management

This chapter begins by providing a fundamental background to the field of Cloud resource management in Section **??**, then addresses several areas of current research interest.

Bibliography