Chapter 1

Introduction

1.1 Problem Statement

The advent of Cloud computing has completely reformed the software industry [**?**]. In Cloud customers' perspective, applications deployed on a Cloud can have unlimited scalability without upfront investment. Beyond that, Cloud offers a *pay-as-you-go* policy which allows Cloud customers to pay the minimum rental of resources on a basis. These two advantages make Cloud computing an attractive option. Virtualization [**?**] is the core technology that not only enables the elastic management of Cloud resource but also can be used to improve the utilization and reduce energy consumption. It maps a physical machine's system resource - including processors, memory, and other devices - into isolated units called *Virtual Machines (VMs)* which allows multiple operating system to run on. In essence, virtualization add an extra layer of software called *Virtual Machine Monitor (VMM)* or *hypervisors* that can deploy, release and migrate VMs at runtime. Numerous VMMs have been designed for x86 commodity machines such as Xen [**?**], Kvm [**?**], and VMware ESX server [**?**].

Server consolidation is a strategy for improving utilization of Cloud resources [**?**]. It uses a *live VM migration techinique* [**?**] to concentrate VMs into fewer physical servers so that a datacenter can accommodate more applications and idle servers can be turned off to save energy. However, Cloud datacenter is a highly dynamic environment with application demand fluctruation, VMs arrival and release. Maintaining a high level of server utilization is a continuous process with different server consolidation methods. Mainly, there are two types of server consolidation: static and dynamic. Static method is often treated as an offline approach and it is applied in a periodical manner where a batch of VMs are allocated to a set of servers. Dynamic method is an online approach, where a single VM needs to be allocated to a set of servers. The overall goal for server consolidation is to maximize the utilization of servers as well as minimize the number of migration.

In the previous decade, Cloud providers who were focusing on ensuring quality of service, has quickly expanded their infrastructures into a large scale. As a result, the average utilization of servers is as low as 20%, according to [**?**]'s observation of google's datacenter, hence, the energy was largely wasted.

Despite the energy consumption by none information and communication technologies (ICT) equipments such as cooling and airing systems, the energy can be derived from two aspects, first, the hardwares of servers contribute a static consumption. Second, the usage of computing, storage and network resources cause a dynamic consumption. Therefore, improving the efficiency of resources are also two folds, minimizing the static part and delievering more performance proportional to the dynamic workload.

These are the inituition behind server consolidation. Server consolidation improves the utilization of resources by concentrates workloads in a few servers so that others can be

turned off or put into sleep to save energy. Therefore, it achieves reduction of static energy consumption as well as improving the utilization of resources. The consolidation can be done with the help of virtual machine (VM), which can be easily transport from one physical machine (PM) to another. However, as the scale of reallocation of VMs become large and various quality of services (QoS) requirements have to be considered, an efficient automatic approach is an urgent and necessary need.

Bibliography