Abstract

Modern data centers integrate a lot of computer and electronic devices. However, some reports state that the mean usage of a typical data center is around 50 % of its peak capacity, and the mean usage of each server is between 10 % and 50 %. A lot of energy is destined to power on computer hardware that most of the time remains idle. Therefore, it would be possible to save energy simply by powering off those parts from the data center that are not actually used, and powering them on again as they are needed.

Most data centers have computing clusters that are used for intensive computing, recently evolving towards an on-premises Cloud service model. Despite the use of low consuming components, higher energy savings can be achieved by dynamically adapting the system to the actual workload. The main approach in this case is the usage of energy saving criteria for scheduling the jobs or the virtual machines into the working nodes. The aim is to power off idle servers automatically. But it is necessary to schedule the power management of the servers in order to minimize the impact on the end users and their applications.

The objective of this thesis is the elastic and efficient management of cluster infrastructures, with the aim of reducing the costs associated to idle components. This objective is addressed by automating the power management of the working nodes in a computing cluster, and also proactive stimulating the load distribution to achieve idle resources that could be powered off by means of memory overcommitment and live migration of virtual machines. Moreover, this automation is of interest for virtual clusters, as they also suffer from the same problems. While in physical clusters idle working nodes waste energy, in the case of virtual clusters that are built from virtual machines, the idle working nodes can waste money in commercial Clouds or computational resources in an on-premises Cloud.