Title: Efficient cloud computing system operation strategies in heterogeneous perspective

Abstract:

Cloud computing systems have emerged as a new paradigm of computing systems by providing on demand based services which utilize large size computing resources. Service providers offer Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) to users depending on their demand and users pay only for the user resources. The Cloud system has become a successful business model and is expanding its scope through collaboration with various applications such as big data processing, Internet of Things (IoT), robotics, and 5G networks.

Cloud computing systems are composed of large numbers of computing, network, and storage devices across the geographically distributed area and   multiple tenants employ the cloud systems simultaneously with heterogeneous resource requirements. Thus, efficient operation of cloud computing systems is extremely difficult for service providers. In order to maximize service providers' profit, the cloud systems should be able to serve large numbers of tenants while minimizing the OPerational EXpenditure (OPEX). For serving as many tenants as possible tenants using limited resources, the service providers should implement efficient resource allocation for users' requirements. At the same time, cloud infrastructure consumes a significant amount of energy. According to recent disclosures, Google data centers consumed nearly 300 million watts and Facebook's data centers consumed 60 million watts. Explosive traffic demand for data centers will keep increasing because of expansion of mobile and cloud traffic requirements. If service providers do not develop efficient ways for energy management in their infrastructures, this will cause significant power consumption in running their cloud infrastructures.

In this thesis, we consider optimal datasets allocation in distributed cloud computing systems. Our objective is to minimize processing time and cost. Processing time includes virtual machine processing time, communication time, and data transfer time. In distributed Cloud systems, communication time and data transfer time are important component of processing time because data centers are distributed geographically. If we place data sets far from each other, this increases the communication and data transfer time. The cost objective includes virtual machine cost, communication cost, and data transfer cost. Cloud service providers charge for virtual machine usage according to usage time of virtual machine. Communication cost and transfer cost are charged based on transmission speed of data and data set size.

Also, this thesis proposes an adaptive data center activation model that consolidates adaptive activation of switches and hosts simultaneously integrated with a statistical request prediction algorithm. The learning algorithm predicts user requests in predetermined interval by using a cyclic window learning algorithm. Then the data center activates an optimal number of switches and hosts in order to minimize power consumption that is based on prediction. We designed an adaptive data center activation model by using a cognitive cycle composed of three steps: data collection, prediction, and activation.

Network Function Virtualization (NFV) emerged as a game changer in network market for efficient operation of the network infrastructure. Since NFV transforms the dedicated physical devices designed for specific network function to software-based Virtual Machines (VMs), the network operators expect to reduce a significant Capital Expenditure (CAPEX) and Operational Expenditure (OPEX). Softwarized VMs can be implemented on any commodity servers, so network operators can design flexible and scalable network architecture through efficient VM placement and migration algorithms. In this thesis, we study a joint problem of Virtualized Network Function (VNF) resource allocation and NFV-Service Chain (NFV-SC) placement problem in Software Defined Network (SDN) based hyper-scale distributed cloud computing infrastructure. The objective of the problem is minimizing the power consumption of the infrastructure while enforcing Service Level Agreement (SLA) of users.

In this thesis, we propose efficient cloud infrastructure management strategies from a single data center point of view to hyper-scale distributed cloud computing infrastructure for profitable cloud system operation. The management schemes are proposed with various objectives such as Quality of Service (Qos), performance, latency, and power consumption. We use efficient mathematical modeling strategies such as Linear Programming (LP), Mixed Integer Linear Programming (MILP), Mixed Integer Non-linear Programming(MINP), convex programming, queuing theory, and probabilistic modeling strategies and prove the efficiency of the proposed strategies through various simulations.