COMPARATIVE STUDY ON CLOUD OPTIMIZED RESOURCE AND PREDICTION USING MACHINE LEARNING ALGORITHM

PRASADU PEDDI
Research Scholar, Sathyabama University, Chennai, India.

Dr. SIVABALAN ARUMUGAM
Professor, Dept of CSE, Sathyabama University, Chennai, India.

ABSTRACT:

Conventionally, the resource allocation is formulated as an optimization problem and solved online with instant scenario information. Since most resource allocation problems are not convex, the optimal solutions are complicated to obtain in real time. Therefore, the conventional methods of resource allocation are facing significant difficulties to meet the ever-increasing QoS requirements of users with scarce radio resource. Assisted by cloud computing, a massive amount of historical data on scenarios can collect for extracting similarities among scenarios using machine learning. Moreover, optimal or near-optimal solutions of historical scenarios can be searched offline and stored in advance. When the measured data of the current scenario arrives, the current scenario is compared with historical scenarios to find the most similar one. Then, the optimal or near-optimal solution in the most similar historical scenario is adapted to allocate the radio resources for the current scenario. An example of beam allocation in multi-user massive multiple-input-multiple-output (MIMO) systems shows that the proposed machine-learning based resource allocation outperforms conventional methods.

***Keywords**: QOS, Machine learning, TARA.*

INTRODUCTION:

Resource provisioning and resource management are vivid fields of research in Cloud computing, resulting in a large number of according solutions [2]. Very often, the focus of resource management is on Quality of Service (QoS)-aware provisioning under given cost constraints, or under another set of rules [26]. Apart from a large number of general solutions for the Software-as-a-Service (SaaS) [11], Platform-as-a-Service (PaaS) [1], and Infrastructure-as-a-Service (IaaS) [9] Cloud service models, more specific solutions exist, e.g., for the execution of scientific workflows [9], business processes [14], or data processing [12] in the Cloud. Resource provisioning involves dynamic allocation by scaling resources up and down depending on the current and future demand. While the individual aims of Cloud resource provisioning solutions differ quite a lot, e.g., taking into account QoS and Service Level Agreements (SLAs) [6], the necessary approach is rather homogeneous: The goal is to distribute task requests onto Cloud-based computational resources, e.g., Virtual Machines (VMs) or containers [15]. Aside from deciding when to scale up or down, a common challenge is to distribute computational tasks across available computational resources. Resource provisioning approaches can be categorized into predictive and reactive strategies [8]. Reactive approaches measure a system's state, e.g., the utilization of a VM, consider current task requests, and take according to actions. In contrast, predictive approaches aim to predict the future behavior of the system. For instance, based on the number of user requests or data packages to be processed, the necessary amount of resources for a future period is calculated. Subsequently, Cloud resources are leased (or released) based on the predicted resource requirements, and tasks distributed among

these resources. These predictive approaches can lead to better resource efficiency and overall response time for the system, since they can adapt to system load in advance, instead of merely reacting in an ad hoc manner [6]. Despite the different goals of the different predictive resource provisioning strategies presented in the literature, one common requirement of all of these approaches is a precise prediction about how many resources are needed to execute the computational tasks. Despite this fundamental requirement, to the best of our knowledge, surprisingly little research has been done so far in the field of prediction mechanisms for Cloud resource utilization. In order to address this challenge for efficient resource provisioning, we present a generic approach to predict Cloud resource utilization and task duration. We apply machine learning approaches to predict these values on a per-task level, based on historical task execution data.

RESOURCE ALLOCATION & ITS SIGNIFICANCE

Resource allocation is the scheduling of the available resources and available activities required by those activities while taking into consideration both the resource availability and the project time. Resource provisioning and allocation solve that problem by allowing the service providers to manage the resources for each request for a resource. Resource Allocation Strategy (RAS) is all about the number of activities for allocating and utilizing scarce resources within the limit of cloud environment to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job From the perspective of a cloud provider, predicting the dynamic nature of users, user demands, and application demands are impractical. For the cloud users, the number of tasks of job needs to complete on time with minimal cost. Hence due to limited resources, resource heterogeneity, environmental necessities, locality restrictions and dynamic nature of resource demand, we need an efficient resource allocation system that suits cloud environments. Cloud resources consist of virtual resources. The physical resources shared across multiple computer quests through virtualisation and provisioning. The virtualised resources described through a set of parameters detailing the processing, memory and disk needs. Mapping virtualised resources can do provisioning of the cloud to physical ones. The software and hardware resources allocated to the cloud applications on-demand basis [1].

MACHINE LEARNING APPROACH

For predicting resource utilisation in the Cloud, we propose the employment of techniques from the field of machine learning. The intuition behind this is that tasks executed on Cloud-based computational resources often do not scale linearly with the cardinality of their input; therefore, the application of simple linear regression models is not sufficient. Furthermore, a task might take a vector of input data instead of a single item, making it necessary to perform multiple linear regression. Therefore, we propose to use machine learning in order to create prediction models from historical data, i.e., past task executions, and extract a model for obtaining future predictions.

Fundamentals of Machine Learning

Machine learning evolved from the study of pattern recognition and aimed at giving computers the ability to decide problems without being explicitly programmed. Among the most popular machine learning approaches are Artificial Neural Network (ANN) models, which are inspired by biological neurons [4]. Such a network consists of artificial neurons, which have a certain number of inputs and an activation function. Other neurons can use its output as input. While choosing a machine learning model, we took into consideration the fair amount of research in the various areas. Compared to models suitable for classification like Bayesian classification [4] or Fisher's linear discriminant [2] as well as regression models like Support Vector Machines, feed-forward ANN models with error backpropagation are well-suited for regression and provide efficient means of statistical pattern recognition [4, 11]. When the actual output for the given input vector is known, an error between the calculated and the actual output calculated and used for training the model. In our case, this happens through back propagation [11], which enables us to use compact models with sufficient generalisation performance [4]. Machine learning distinguishes between offline and online learning: Offline learning occurs when all instances are presented simultaneously, while in the case of online learning, problem instances presented one at a time. In our instance, we perform offline learning, as training the ANN model involves processing large amounts of data, thus requiring a high amount of processing power. Furthermore, we employ supervised learning, i.e., we assume that for each instance of training data.

Machine Learning Framework

In the machine learning framework, a machine learning algorithm is adopted to build a predictive mode nine no labels, and the goal instead is to organise the data and find hidden structures in unlabeled data. Most machine learning algorithms supervised. In the following, we will discuss how to apply the supervised learning to solve the resource allocation problem. A. Feature Selection In machine learning, feature selection, also known as attribute selection, is the process of selecting a subset of relevant attributes in historical data to form a feature vector for building predictive models. The selection of an appropriate feature vector is critical due to the phenomenon is known as "the curse of dimensionality" [9]. That is, each dimension that added to the feature vector requires exponentially increasing data in the training set, which usually results in practical significant performance degradation. Therefore, it is necessary to find a low dimension of feature vectors that captures the essence of resource allocation in practical scenarios. In order to reduce the dimensionality of feature vectors, only valuable information for the resource allocation can select as features. After modelling the resource allocation as the optimisation problem (1), all valuable information is included in the parameter vector a. Observing the elements of a, it can found that they can further divide into two categories: time-variant (dynamic) or time-invariant (static). Some elements are constants and thus labelled as time-invariant

parameters, such as subcarrier number, maximum transmit power, and antenna number. Other elements that change quickly and are required to measured and feedback all the time for making decisions of the resource allocation labelled as time-variant parameters, such as user number, CSI of all users, and interference levels. As the time-invariant parameters keep unchanged, in order to minimise the dimension of the feature vectors, only the time-variant parameters can be considered to be features. Moreover, some time-variant parameters cannot be selected as features since it may be redundant in the presence of another relevant feature with which it is strongly correlated. In short, an individual feature vector specifies a unique scenario for resource allocation. However, it should noted that the feature selection is a process of trial and error, which can 10 be time-consuming and costly especially with huge datasets. B. Solutions of Optimization Problems To facilitate the application of supervised learning, the solution of resource allocation problem specified by each training feature vector should obtain in advance. Then, each training feature vector associated with its solution. According to the associated solutions, all feature vectors labelled into multiple classes. More specifically, all training feature vectors with the same solution placed with the same class label, indexed by a nonnegative integer. In other words, each class associated with its unique solution. The class label information of all training feature vectors will be used to build a predictive model. In practice, the measured data of a real-time scenario s selected as a new feature vector. Then the predictive model will predict the class for the new feature vector and output the associated solution of the predicted class, i.e., how to allocate the radio resource for the real-time scenario. If too many training feature vectors associated with low-performance solutions, the built predictive model cannot supply high-performance solutions for possible resource allocation. Therefore, finding optimal or near-optimal solutions of all training feature vectors is crucial for building a high-performance predictive model. In the resource allocation problem (1), all elements in the vector x are used to describe how to allocate the radio resources. Mathematically, the allocation of many radio resources can describe by integer variables, such as subcarriers, timeslots, modulation and coding schemes. Intuitively, the transmit power level can be adjusted arbitrarily between the maximum transmit power and zero. It seems that only a continuous variable can be used to describe the transmit power allocation. However, in order to simplify the system complexity, the transmitter in practical systems are usually allowed to transmit signals with only a few prefixed power levels. Therefore, most practical resource allocation issues can be modeled as an integer optimization problem. When the number of integer variables in an integer optimization problem is very small, the optimal solution can be found by exhaustive search

Resource Allocation Strategies & Algorithm

Recently many resource allocation strategies have come up in the literature of the cloud computing environment as this technology has started maturing. A number of Researcher communities around the

world have proposed and implemented several types of resource allocation. Some of the strategies for resource allocation in cloud computing environment are discussed here briefly.

Topology-Aware Resource Allocation (TARA) Different types of resource allocation strategies are proposed in the cloud. The author mentioned in [2] the architecture for resource allocation in Infrastructure-as-a-Service (IaaS) based cloud systems. Current Infrastructure-as-a-Services of cloud providers are usually unaware of the hosted application's requirements and therefore allocate resources independently of its needs, which can significantly impact performance for distributed data-intensive applications An architecture which adopts a "what if" methodology to guide allocation decisions taken by the IaaS is proposed to address this resource allocation problem. The architecture uses a prediction engine with a lightweight simulator to estimate the performance of given resource allocation and an algorithm to find an optimized solution in the large search space. Results showed that Topology Aware Resource Allocation reduced the job completion time of these applications by up to 59% when compared to application-independent allocation policies.

Linear Scheduling Strategy for Resource Allocation Considering the processing time, resource utilization based on CPU usage, memory usage and throughput, the cloud environment with the service node to control all clients request, could provide maximum service to all clients [3]. Scheduling the resource and tasks separately involves more waiting time and response time. A scheduling algorithm named as Linear Scheduling for Tasks and Resources (LSTR) is designed, which performs tasks and resources scheduling respectively. Here, a server node is used to establish the IaaS cloud environment and KVM/Xen virtualization along with LSTR scheduling to allocate resources which maximize the system throughput and resource utilization.

Dynamic Resource Allocation for Parallel Data Processing Dynamic Resource Allocation for Efficient Parallel data processing [4] introduces a new processing framework explicitly designed for cloud environments called Nephele Most notably, Nephele is the first data processing framework to include the possibility of dynamically allocating/de-allocating different compute resources from a cloud in its scheduling and during job execution. Particular tasks of a processing job can be assigned to different types of virtual machines which are automatically instantiated and terminated during the job execution.

COMPARATIVE STUDY

An optimal RAS should avoid the following criteria as follows:

- Resource Contention - Resource contention arises when two applications try to access the same resource at the same time.
- The scarcity of Resource - Scarcity of resource arises when there are limited resources and the demand for resources is high.
- Resource Fragmentation - Resource fragmentation arises when the

resources are isolated. There would be enough resources but cannot allocate it to the needed application due to fragmentation into small entities.

- Over Provisioning - Over provisioning arises when the application gets surplus resources than the demanded one. Under Provisioning - Under the provisioning of resources occurs when the application is assigned with fewer numbers of resources than it demanded.

CONCLUSION

Cloud computing technology is increasingly being used in enterprises and business markets. A review shows that dynamic resource allocation is the growing need for cloud providers for more number of users and with the less response time. In the cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper summarizes the main types of RAS and its impacts on the cloud system. Some of the strategies discussed above mainly focus on memory resources but are lacking in other factors. Hence this survey paper will hopefully motivate future researchers to come up with smarter and secured optimal resource allocation algorithms and framework to strengthen the cloud computing paradigm.

REFERENCES

1. V. Vinothina, Dr. R. Shridaran, and Dr. Padmavathi Ganpathi, A survey on resource allocation strategies in cloud computing, International Journal of Advanced Computer Science and Applications, 3(6):97--104, 2012.

2. Gunho Lee, Niraj Tolia, Parthasarathy Ranganathan, and Randy H. Katz, Topology-aware resource allocation for data-intensive workloads, ACM SIGCOMM Computer Communication Review, 41(1):120--124, 2011.

3. Abirami S.P. and Shalini Ramanathan, Linear scheduling strategy for resource allocation in a cloud environment, International Journal on Cloud Computing: Services and Architecture(IJCCSA), 2(1):9--17, 2012.

4. Daniel Warneke and Odej Kao, Exploiting dynamic resource allocation for efficient parallel data processing in the cloud, IEEE Transactions On Parallel And Distributed Systems, 2011.

5. Atsuo Inomata, Taiki Morikawa, Minoru Ikebe and Md. Mizan Rahman, Proposal and Evaluation of DynamicResource Allocation Method Based on the Load Of VMs on IaaS, IEEE, 2010.

6. Dorian Minarolli and Bernd Freisleben, Utility-based Resource Allocations for virtual machines in cloud computing, IEEE, 2011.

7. Jiyani, Adaptive resource allocation for preemptable jobs in cloud systems, IEEE, 2010.

8. Bo An, Victor Lesser, David Irwin and Michael Zink, Automated Negotiation with decommitment for Dynamic Resource Allocation in Cloud Computing, Conference at the University of Massachusetts,Amherst, USA.

9. Gihun Jung and Kwang Mong Sim, Location-Aware DynamicResource Allocation Model for Cloud Computing Environment,International Conference on Information and Computer Applications(ICICA), IACSIT Press, Singapore, 2012.

10. Chandrashekhar S. Pawar and R.B. Wagh, A review of resource allocation policies in cloud computing, World Journal of Science and Technology, 2(3):165-167, 2012.

11. Sandeep Tayal, Tasks Scheduling Optimization for the Cloud Computing systems, International Journal of Advanced Engineering Sciences and Technologies (IJAEST), 5(2): 111 - 115, 2011.

12. Ronak Patel, Sanjay Patel, "Survey on Resource Allocation Strategies in Cloud Computing", International Journal of Engineering and Research Technology, Vol 2, Issue 2, February 2013.

13. K Prasanna Kumar, S Arun Kumar, Dr Jagadeeshan, "Effective Load Balancing for Dynamic Resource Allocation in Cloud Computing", International Journal of Innovative Research in Computer and Communication Engineering,Vol 2 , Issue 3,March 2013.

14. Yang Wt Al,"A Profile-Based Approach to Just in Time Scalability for Cloud Applications", IEEE International Conference, 2009.

15. Peter Mell, Timothy Grance," The NIST Definition of Cloud Computing (Draft)", Computer Security Division, Information Technology Laboratory, 2011