The feasibility and properties of dividing virtual machine resources using the virtual machine cluster as the unit in cloud computing

Zhiping Peng[1], Bo Xu[1, 2,*], Antonio Marcel Gates[4], Delong Cui[1], and Weiwei Lin[3]
[1] Department of Computer Science and Technology, Guangdong University of Petrochemical Technology
Maoming, Guangdong 525000 - CHINA
[e-mail: xubo807127940@163.com]
[2] School of Software Engineering, South China University of Technology
Guangzhou, Guangdong 510006 - CHINA
[3] School of Computer Science and Engineering, South China University of Technology
Guangzhou, Guangdong 510006 - CHINA
[4] Hawaii Pacific University
Honolulu, Hawaii 96813 - USA
*Corresponding author: Bo Xu

Received February 13, 2015; revised May 18, 2015; accepted June 8, 2015; published July 31, 2015

Abstract

In the dynamic cloud computing environment, to ensure, under the terms of service-level agreements, the maximum efficiency of resource utilization, it is necessary to investigate the online dynamic management of virtual machine resources and their operational application systems/components. In this study, the feasibility and properties of the division of virtual machine resources on the cloud platform, using the virtual machine cluster as the management unit, are investigated. First, the definitions of virtual machine clusters are compared, and our own definitions are presented. Then, the feasibility of division using the virtual machine cluster as the management unit is described, and the isomorphism and reconfigurability of the clusters are proven. Lastly, from the perspectives of clustering and cluster segmentation, the dynamics of virtual machines are described and experimentally compared. This study aims to provide novel methods and approaches to the optimization management of virtual machine resources and the optimization configuration of the parameters of virtual machine resources and their application systems/components in large-scale cloud computing environments.

Keywords: Virtual resources, virtual machine cluster, graph cuts theory, graph theory, cloud computing

This research was supported by research grants from the National Natural Science Foundation of China (No.61272382, 61402183). and grants from Science and Technology Planning Project of Guangdong Province, China (No.2013B010401005, 2013B010401024)

http://dx.doi.org/10.3837/tiis.2015.07.018 **ISSN : 1976-7277**

1. Introduction

Virtualization technology has been one of the driving forces behind the rapid development of cloud computing in recent years [1]. It abstracts the basal architecture, such as the physical resources in cloud computing, so that the differences and compatibilities of different types of equipment are transparent to the upper applications, which makes the unified management of the vastly different hardware resources underlying the cloud possible. In addition, virtualization technology simplifies the programming of applications, which allows developers to focus solely on the business logic and ignore the supply and configuration of the basal resources. Through virtualization technology, a single physical server can support multiple virtual machines to run multiple operating systems and applications. These applications reside on their respective virtual machines and constitute a certain type of isolation such that the collapse of an application will not affect the operation of the other applications. Lastly, the ease of creating a virtual machine makes it possible for the application to have more virtual machines for fault tolerance and disaster recovery so that the virtual machine's reliability and applicability are improved. However, the introduction of virtualization technology has not reduced the complexity of the management of the relevant configurations in cloud computing. In fact, the operation of multiple virtual machines on the same physical computing infrastructure creates difficulties for the whole management system and raises new challenges [2].

A cloud computing environment is an open and heterogeneous environment in which load, infrastructure, virtual machines, and the development of applications are rapidly changing [3]. From the perspective of both the supply and demand of cloud services, resource utilization and service-level agreements are the two fundamental issues that interest and concern both cloud providers and cloud users [4]. In the dynamic cloud computing environment, to ensure, under the terms of service-level agreements, the maximum resource utilization, it is necessary to investigate the online dynamic management of the parameters of virtual machine resources and their operational application systems/components [5]. With the development of cloud computing, the management of virtual machines in the cloud environment has attracted considerable recent attention [6, 7]. To meet the needs of a variety of applications by users, researchers have extensively investigated the mechanisms of virtual machines, and the focus of studies has shifted from single virtual machines to virtual machine clusters [8]. In the cloud computing environment, the providers of applications typically deploy some virtual machines from their leased physical resources to provide services to end-users, and these virtual machines form a virtual machine cluster, a collection of multiple virtual machines that belong to a certain user's application with communication requirements and deployment constraints [9].

Graph theory, which is abstract algebra that studies the relationships among specific entities, plays a very important role in computer science. It applies abstraction to practical problems to construct models, form graphs with vertices and edges, and find solutions using the basic algorithms of graph theory [10]. In 1736, graph theory was first proposed by Euler, and ever since, it has been gradually improved by various scholars. Graph theory is a branch of mathematics that mainly uses graphs as the research object to study mathematical theories and methods of graphing that consist of vertices and edges [11]. This study employs graph theory to investigate the formalization, feasibility, and related properties of division on the cloud platform, using the virtual machine cluster as the management unit.

In this paper, the feasibility and properties of the division of virtual machine resources on the cloud platform, using the virtual machine cluster as the management unit, are investigated. We present existing definitions of virtual machine clusters, and propose our own definition with comparison. The isomorphism and reconfigurebility of a cluster are discussed and proven, and evaluated by experiments. The remaining parts of this paper are organized as follows. In Section 2, we discuss recent related work. In Section 3, the definitions of virtual machine clusters are compared, and our own definitions are presented. In Section 4, the feasibility of division using the virtual machine cluster as the management unit is described, and the isomorphism and reconfigurability of the clusters are proven. Lastly, from the perspectives of clustering and cluster segmentation, the dynamics of virtual machines are described and experimentally compared, and our conclusions in Section 5.

2. Related Work

In recent years, many achievements in virtualization technology have been made in the field of cloud computing, leading to significant developments in cloud computing technology. Currently, studies on virtual machine clusters containing multiple virtual machine resources rather than single virtual machine resources have attracted considerable attention [12-14]. Wang et al. argued that the deployment of a series of interlinked virtual machines in the form of cluster makes it easier to achieve the overall efficiency of cloud computing and cloud services; under such a rationale, they created clusters from the virtual machine resources and physical hosts and realized the deployment of virtual machine clusters through the inter-cluster matching [15]. Yao et al. proposed a deployment algorithm that restricts both the resource and traffic volume to address the issue of virtual machine cluster deployment in the cloud computing environment, which they compared with a greedy algorithm and a single-constraint algorithm, finding that the new algorithm enhanced the efficiency of the communication bandwidth utilization in the system [16]. Xiaohui.Wei proposed a Topology-aware Partial Virtual Cluster Mapping algorithm (TOP-VCM) which is based on sub-graph isomorphism detection. TOP-VCM can fully satisfy the nodes/links requirements in Communication Skeleton to ensure the execution performance with only slight degradation of other trivial nodes/links to significantly reduce the mapping difficulty. Simulation results have shown that TOP-VCM has significantly improved the total revenue, the utilization of physical resources and the performance of mapping algorithm while satisfying the Virtual Cluster requirements [17]. Yumei Huang introduce an innovative snapshot approach for virtual cluster that exploits shared memory pagesamong all the component VMs to reduce the size of produced snapshot and mitigate the I/O bottleneck [18]. Xinkui Zhao. present a novel cluster performance optimization strategy named vClusterOpt. vClusterOpt finds out centralized subgraphs of node graph and choose node with the shortest logical distance as kernel node of the subgraph to reduce inter-machine communication and transmission cost under virtual cluster.Experiments show that an average of 20% performance improvement can get by distance-aware virtual cluster optimization strategy [19]. Wenyu Zhou designed and implemented a load balancing scheme based on dynamic resource allocation policy for virtual machine cluster, It optimize the resource allocation of VMs to achieve global load balancing of virtual machine cluster. Compared to traditional load balancing schemes based on task scheduling, it is application independent and works seamless on VMs hosting different kinds of applications [20]. However, most of these studies have only focused on the deployment of virtual machine clusters. The definition, feasibility, and nature of virtual machine clusters, which are the focus of this study, have rarely been addressed [21, 22].

3. Definition of Virtual Machine Clusters

A virtual machine cluster is currently defined as a collection consisting of multiple virtual machines that are closely connected and typically constitute the deployment unit of a certain application. In the cloud computing environment, application providers typically provide services to end-users by deploying some virtual machines from their leased physical resources, in which the virtual machines form the virtual machine cluster, i.e., the collection of virtual machines that belong to the application of a certain user and have communication requirements and deployment constraints [23, 24]. Although the above-mentioned definition has a certain level of rationality, it does not sufficiently address the competition and the dependencies among the virtual machines, nor does it take into account the impact of the application systems [25]. The definition of virtual machine clusters should consider the competition for resources among the virtual machines in the same physical machine and the interdependence among different components deployed on different physical machines with the application systems of a multi-layer structure; in addition, to improve service quality, the definition of virtual machine clusters also needs to consider the users' virtual machine clusters that are effectively deployed by the providers of the cloud infrastructure so that it meets the reliability requirements and has high-level application performance and a reasonable virtual machine deployment time.

The meaning of the term cluster is a collection. Therefore, the virtual machine cluster is actually a collection consisting of multiple virtual machines. Certainly, for those virtual machines that fall into the same virtual machine cluster, they are more or less connected to each other, or simultaneously, they comply with the conditions for access to the given virtual machine cluster. In cloud computing, the virtual machine resources that access a virtual machine cluster are typically for the service of the same application. The realization of the application generally needs those virtual machine resources to be deployed on the physical server within the cloud; virtual machine clusters that can provide high-quality service for the application are inevitably those that collectively meet expectations regarding resource requirements, bandwidth, etc. Taking these considerations into account, the relevant virtual machine cluster is defined as follows:

Definition 1: A group of related virtual machines on a cloud platform with a multi-layer application system (e.g., Web applications that adopt the Java 2 Platform, Enterprise Edition (J2EE) 3-tier structure) that operates interdependently or competes for resources is designated a virtual machine cluster.

Definition 2: Whether it is a complete single-layer application or one layer of a multi-layer application system, the service that operates on one virtual machine within the cluster is collectively designated an application system/component.

To facilitate understanding, an example is presented here. **Fig. 1** shows the schematic diagram of a virtual machine cluster. It contains 2 physical servers (shown as solid boxes), each of which creates 2 virtual machines (shown as grey circles); there exists a relationship based on competition for resources among the virtual machines on the same physical server (i.e., with the overall resources in the physical server constant, zero-sum relations exist among the virtual machines). There are 2 application systems, both with a double-layer structure (app1 and app2, shown as red dashed boxes), each of which is deployed on the virtual machine located on different servers, and the overall performance of application systems (e.g., response time, throughput, etc.) is dependent on the setting of the component parameters and the interdependence among the different components.

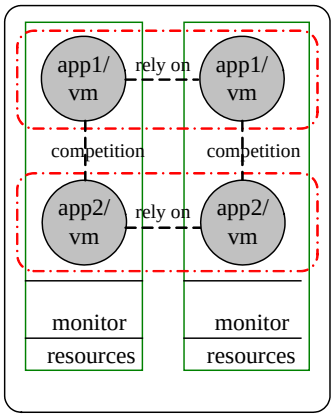

Fig. 1. A virtual machine cluster.

When multiple virtual machines form a virtual machine cluster, the virtual machines within the cluster eventually complete the task through synergies among the machines. In the process of cooperative completion of the task, the virtual machines interactively communicate with each other due to the existence of the competitive and interdependent relations, thus forming a hierarchical network structure. The hierarchical network organisation of a typical virtual machine cluster is shown in **Fig. 2**.

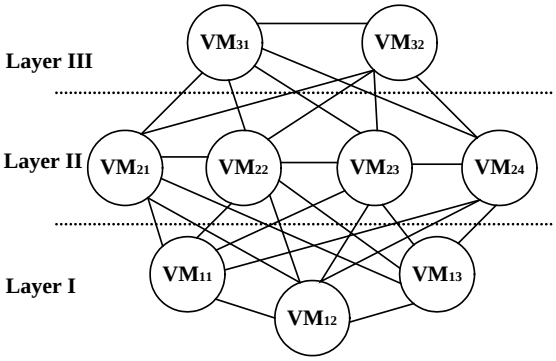

Fig. 2. Hierarchical network structure of a virtual machine cluster.

As shown in **Fig. 2**, the virtual machine cluster is a 3-layer network structure. There are 3, 4, and 2 virtual machine units in the first layer, second layer, and third layer, respectively, and among the virtual machines, there are inter-layer interactions, as well as intra-layer interactions, that ensure the coordination necessary to complete the same task.

4. The Feasibility and Related Properties of Dividing Virtual Machine Resources using the Cluster as the Unit

4.1 Feasibility Analysis

The description of the feasibility of dividing virtual machine resources using the cluster as the management unit is mainly based on 2 aspects. On the one hand, when a user sends a request to the cloud, the cloud allocates the virtual machine resources according to the actual situation of the task. In many cases, to meet the user's demand and complete the task at the highest level of efficiency, it is necessary to gather the virtual machine resources to form a cluster. Gathering the virtual machine resources based on a certain demand to form a virtual machine cluster provides considerable convenience for the subsequent configuration of the hardware resources with the virtual machine cluster and greatly enhances the efficiency of the whole service. Because the existing interdependence or competitive relationships among the virtual machines in the cloud environment require communicative interactions, the issue of communication bandwidth among the virtual machines is important. This type of real-world issue can be abstracted to form graphs with vertices and edges, in which the connection lines among the virtual machines represent the communication needs among the virtual machines, for modelling, and it is feasible to solve the problem using the basic algorithms of graph theory.

On the other hand, in the actual implementation of cloud computing tasks, clustered virtual machines need to be matched based on the physical resources, i.e., allocating virtual machine resources to the appropriate host. One physical host can be allocated to multiple virtual machines, but when the capacity of the physical host cannot meet the needs of a virtual machine cluster, it is necessary to reasonably divide the virtual machine cluster and individually allocate the resultant segments to different hosts, which creates the issue of cluster segmentation. Dividing the cluster into individual sub-clusters and selecting the deployment of the target host according to the constraint conditions is necessary to enhance the efficiency of resource utilisation and reduce the system's communication bandwidth requirements. Virtual machine cluster segmentation is essentially a division problem. The use of graph theory to solve division problems has achieved effective results [26, 27]. Virtual machine cluster segmentation is relatively more complex; the minimum cut algorithm in graph theory can achieve optimal partitioning with graphs that have constraint conditions, properly meeting the context-specific requirements of the virtual machine cluster problem.

4.2 Cluster Isomorphism

Fig. 3 shows a virtual machine cluster that consists of multiple virtual machines. Each of the virtual machines requires communication interactions due to the interdependence or competitive relations among the virtual machines, and the connection lines among the virtual machines represent these communication needs, resulting in a communication bandwidth problem among the virtual machines. Through modelling, the communication bandwidth among the virtual machines is constrained and abstracted to the undirected edges of the graph, generating the undigraph model of the virtual machines from **Fig. 3** to **Fig. 4**, which is subject to further analysis.

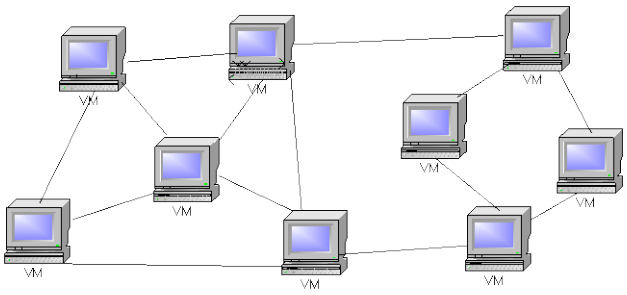

Fig. 3. The structure of a virtual machine cluster.

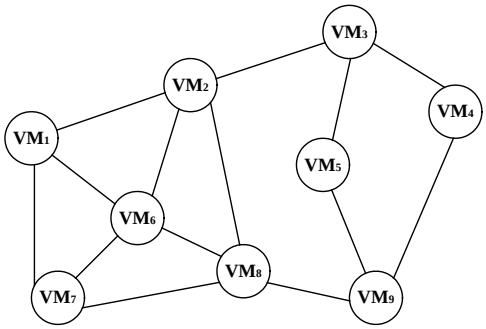

Fig. 4. The graph structure corresponding to a virtual machine cluster

Definition 3: The virtual machine cluster corresponding to the undirected graph is $G_{VMC} = (V_{VMC}, E_{VMC})$, in which V_{VMC} is the set of vertices of the virtual machine cluster and $E_{VMC} = \{e(n_i, n_j) \mid n_i, n_j \in V_{VMC}\}$ is a set of undirected edges.

Theorem 1 (cluster isomorphism). Assume that $G_1 = (V, E)$ and $G_2 = (V', E')$ are two graphs corresponding to virtual machine clusters (VMC_1) and virtual machine clusters (VMC_2), respectively; if there is one-to-one $\varphi : V \to V'$ such that any 2 given vertices u and v $(u, v) \in E$, if and only if $(\varphi(u), \varphi(v)) \in E'$ and (u, v), and $(\varphi(u), \varphi(v))$ have the same multiplicity, G_1 and G_2 are defined as isomorphic, denoted as $VMC_1 \cong VMC_2$.

Corollary 1 (the necessary condition for cluster isomorphism): when VMC_1 and VMC_2 are isomorphic, they exhibit the following properties:

1) $|V| = |V'|$, $|E| = |E'|$.
2) The degrees of v and $\varphi(v)$ are equal.
3) The number of the fixed point with the same degree is equal.
4) $(u, v) \in E$, if and only $(\varphi(u), \varphi(v)) \in E'$.
5) If 4) holds, then (u, v) and $(\varphi(u), \varphi(v))$ have the same multiplicity.

Theorem 2 (cluster reconfiguration). A cluster corresponding to graph $G(n \geq 2)$ is defined as reconfigurable if *G* can be uniquely identified (in the context of isomorphism) by its sub-graphs $G - v_i (1 \leq i \leq n)$. The cluster corresponding to graph $G(n \geq 2)$ is also reconfigurable.

Proof: Suppose *G* is a graph with an order $n \geq 2$ and m number of edges, and set $V(G) = \{v_1, v_2, ...v_n\}$ and *e* as an edge of *G*. The edge *e* then appears in each of the sub-graphs $G - v_i (2 \leq i \leq n)$, but not in $G - v_1$. Set the number of edges of $G - v_i$ to be $m_i (1 \leq i \leq n)$; then, in the sum formula of $\sum_{i=1}^{n} m_i$, each of the edges are counted $n-2$ times; i.e., $\sum_{i=1}^{n} m_i = m(n-2)$, $m = \frac{\sum_{i=1}^{n} m_i}{n-2}$. This proves that the number of edges in each of the graphs is determinable. When the number of edges of the graph *G* is *m* and then the vertex v_i is removed from the graph *G*, the number of edges of the resultant $G - v_i$ is $m_i (1 \leq i \leq n)$. Therefore, $\deg v_i = m - m_i$. Accordingly, $m - m_1$, $m - m_2$, …, $m - m_n$ are a degree sequence of the graph *G*. At this point, from the sub-graphs of $G - v_i$ derived from the graph G, not only the order of *G*, the number of edges, and the degree sequence but also the graph itself can be uniquely determined. Thus, it is proven that G ($n \geq 2$) is reconfigurable, and its corresponding virtual machine cluster is also reconfigurable.

Because there is no competition for resources among the clusters, the performance of application systems/components does not constitute interdependence; once one cluster achieves the optimal configuration strategy through the machine learning methods, it can migrate the knowledge to other clusters with similar applications, which is an effective way to simplify complex learning problems. The isomorphic cluster provides good and effective solutions to the foregoing problems, which is also the objective of the study on cluster isomorphism and its reconfigurable nature.

4.3 Cluster Dynamics

Cluster dynamics mainly refer to the 2 steps of clustering and cluster segmentation, which are asynchronous and dynamic, that the virtual machine cluster must undergo to ensure improved resource utilisation efficiency under the terms of service-level agreements.

(1) Clustering

Virtual machine resources are clustered to form a virtual machine cluster that provides considerable convenience for the subsequent configuration of virtual machine resources with the hardware resources and greatly enhances the efficiency of the whole operational service. In this study, 2 types of determination conditions are adopted for the clustering of virtual machines:

(1) The clustered virtual machine resources should be the application service for a certain user or a certain application system/component, and therefore, they should have identical or similar requirements concerning hardware resources.

(2) The bandwidth requirements for communication among the clustered virtual machine resources should be relatively similar.

H_{CPU} denotes the CPU hardware resource requirements of the virtual machines; H_{Mem}

stands for the memory hardware resource requirements of the virtual machines; and H_{Hard} represents the hard disk hardware resource requirements of the virtual machines. The determination function $S_{i,j}(H)$ of the difference in the hardware resource requirements between the virtual machine *i* and the virtual machine *j* can be constructed and is presented in Equation (1):

$$S_{i,j}(H) = \theta_1 \left\| H^i_{CPU} - H^j_{CPU} \right\| + \theta_2 \left\| H^i_{Mem} - H^j_{Mem} \right\| + \theta_3 \left\| H^i_{Hard} - H^j_{Hard} \right\| \quad (1)$$

where θ_1, θ_2, and θ_3 are the respective weights of the differences in the hardware resources between the i^{th} and j^{th} virtual machines, and B_i and B_j are the respective communication bandwidth requirements in the communication processes of the i^{th} and j^{th} virtual machines. The determination function $S_{i,j}(B)$ of the difference in bandwidth resource requirements between the 2 virtual machines is presented in Equation (2):

$$S_{i,j}(B) = \left\| B_i - B_j \right\| \quad (2)$$

Based on the similarity determination function formed according to the constraint conditions on 2 two types of resources, an overall determination function $S_{i,j}$ for the clustering of virtual machines is constructed and presented in Equation (3):

$$S_{i,j} = \omega_1 S_{i,j}(H) + \omega_2 S_{i,j}(B) \quad (3)$$

where ω_1 and ω_2 are the weights of the respective similarity determination functions of the 2 types of constraint resources in the overall determination function. Through comparison of the computing result using the determination function and the preset threshold value, virtual machines with similar attributes can be clustered to construct a virtual machine cluster via the appropriate communication connection.

(2) Cluster segmentation

When a user sends a request to the cloud, the cloud configures virtual machine resources for the user according to the actual situation of the task. In many cases, to meet the needs of users and complete the task at the highest level of efficiency, it is necessary to cluster the virtual machine resources into a cluster using a clustering method such as the one described in this study. In the actual implementation of a cloud computing task, the clustered virtual machines need to match based on the physical resources, i.e., allocating virtual machine resources to the appropriate hosts. One physical host can be assigned multiple virtual machines; however, when the performance of a physical host cannot meet the needs of a virtual machine cluster, it is necessary to reasonably divide the virtual machine cluster, which creates the cluster segmentation problem. Below, an example is used to describe the process and procedures of cluster segmentation. A given cluster structure that has completed the clustering process is shown in **Fig. 5**.

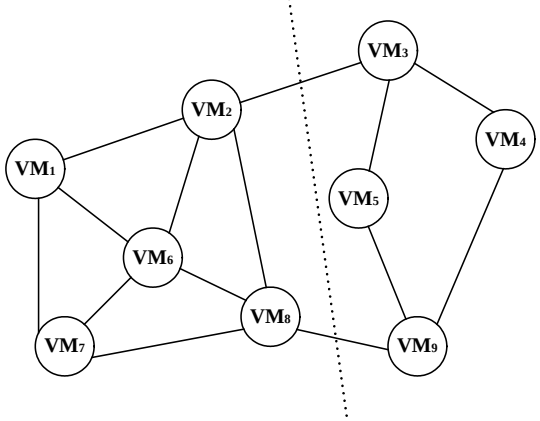

Fig. 5. The cluster segmentation of a virtual machine cluster

This cluster provides services for the same task; however, when configuring the task with the physical host, the task cannot be successively configured in a single physical host most likely because the resource needs of the entire cluster are too large, which creates the cluster segmentation problem for the virtual machine cluster structure. Formally, segmentation in the virtual machine cluster structure and the cutting problem in graph theory are remarkably similar. Based on this observation, graph cuts theory is used to construct a mathematical model of the virtual machine cluster to apply a graph cuts approach to cluster segmentation. For the virtual machine cluster structure in Figure 5, each of the virtual machines is set as the vertex in the cluster network structure and designated VM_i. From Figure 1, it can be seen that the virtual machine cluster contains a total of 9 vertices. The attributes of each virtual machine vertex can be depicted based on its required resources such that the attributes of each virtual machine can be portrayed in the form of a vector, as shown in Equation (4).

$$A_{VM_i} = \{R_{VM_i}^{CPU}, R_{VM_i}^{Mem}, R_{VM_i}^{Hard}, R_{VM_i}^{BW}, R_{VM_i}^{TL}\} \tag{4}$$

The 2-way connection between 2 virtual machines, with its attributes depicted as $B_{VM_{ij}}$, represents the bandwidth resources necessary for the virtual machines to communicate with each other. When configuring the entire virtual machine cluster to a physical host, the upper limit of the resources provided by the physical host needs to be determined and compared with the resource needs of the entire virtual machine cluster. When the resources provided by the physical host are able to meet the requirements of the entire virtual machine cluster, the resources can be deployed directly; otherwise, it is necessary to conduct fragmentation deployment for the virtual machine cluster. The direct deployment must meet the following conditions simultaneously:

$$\begin{cases} R_{PM}^{CPU} > \sum_i R_{VM_i}^{CPU} \\ R_{PM}^{Mem} > \sum_i R_{VM_i}^{Mem} \\ R_{PM}^{Hard} > \sum_i R_{VM_i}^{Hard} \\ R_{PM}^{BW} > \sum_i R_{VM_i}^{BW} + \sum_j B_{VM_{ij}} \end{cases} \tag{5}$$

When it is necessary to segmentation-deploy the virtual machine cluster so that the redeployment is possible, the segmentation process can be performed according to the graph cuts theory. That is, to segment the network in Figure 5 to form a series of sub-networks, each sub-network must match the physical host to complete its configuration. Here, each segmentation is not necessarily required to have the minimum cut, provided that the resultant sub-networks find their matching hosts. In this regard, cluster segmentation is different from the general graph cuts method.

A virtual machine cluster is assumed to originally contain *n* virtual machines. After segmentation, *m* sub-clusters are formed, with each sub-cluster containing $n_1, n_2, \cdots, n_m$ virtual machines and respectively arranged in *m* physical hosts. After arrangement, the result complies with the following:

$$\begin{cases} \forall_k R_{PM}^{CPU} > \sum_{i=1}^{n_k} R_{VM_i}^{CPU} \\ \forall_k R_{PM}^{Mem} > \sum_{i=1}^{n_k} R_{VM_i}^{Mem} \\ \forall_k R_{PM}^{Hard} > \sum_{i=1}^{n_k} R_{VM_i}^{Hard} \\ \forall_k R_{PM}^{BW} > \sum_{i=1}^{n_k} R_{VM_i}^{BW} + \sum_j B_{VM_{ij}} \end{cases} \tag{6}$$

Where R_{PM}^{CPU} represents CPU resources configuration of the physical host, R_{PM}^{Mem} represents memory resources configuration of the physical host, R_{PM}^{Hard} represents hard disk resources configuration of the physical host, R_{PM}^{BW} represents bandwidth resources configuration of the physical host, $R_{VM_i}^{CPU}$ represents CPU resources requirements of the virtual machine, $R_{VM_i}^{Mem}$ represents memory resources requirements of the virtual machine, $R_{VM_i}^{Hard}$ represents hard disk resources requirements of the virtual machine, $R_{VM_i}^{BW}$ represents bandwidth resources requirements of the virtual machine, $B_{VM_{ij}}$ represents virtual machine communication bandwidth requirements between i virtual machine and j virtual machine.

The meaning of the above formula is that a total of n_k virtual machine clusters deployment on the physical host K, which must make physical host in terms of CPU, memory,

hard disk, bandwidth is greater than the sum of the resource requirements of a virtual machine at the same time.

5. Experimental Classification Results and Analysis

To verify the effectiveness of the clustering of virtual machines and the cluster segmentation method, the following experiment is conducted. First, virtual machines with similar attributes are selected from 200 virtual machines according to the needs of users, and they form a virtual machine cluster consisting of 9 virtual machines. The communication relationships among the virtual machines are shown in Figure 1. The hardware resource requirements of the 9 virtual machines are listed in **Table 1**.

Table 1. Hardware resource requirements of the virtual machines.

Machine Number	CPU (MIPS)	REM (GB)	Hard disk (GB)	Bandwidth(MB/s)
Virtual machine 1	400	0.8	10	50
Virtual machine 2	500	0.8	10	60
Virtual machine 3	400	0.8	10	50
Virtual machine 4	300	0.6	15	40
Virtual machine 5	300	0.6	10	40
Virtual machine 6	500	1.0	15	60
Virtual machine 7	400	1.0	15	50
Virtual machine 8	500	1.0	20	60
Virtual machine 9	400	1.0	20	50

Here, we need to explain our bandwidth requirements for each virtual machine, in addition to the bandwidth required for communication with the connecting virtual machines, each of the virtual machines is required to have a backup bandwidth of 20 MB/s. We reserved bandwidth of 20 MB/s is given when the virtual machine deployment to the physical host, could happen isolated from the original cluster, incorporated into new clusters, and the associated virtual machine communication needs have a float at that time. In this experiment, communication among the virtual machines requires a bandwidth of 10 MB/s. If a VM interactions to many other VMs, a larger bandwidth would be sufficient. Similarly, a VM interactions to few VMs may need less bandwidth. For example, virtual machine 2 must communicate with virtual machine 1, virtual machine 3, virtual machine 6,and virtual machine 8; therefore, the total bandwidth requirement of virtual machine 2 is 60 MB/s. virtual machine 4 must communicate with virtual machine 3 and virtual machine 9, therefore, the total bandwidth requirement of virtual machine 4 is 40 MB/s.

In the experiment, 5 physical hosts that allow for the deployment of the virtual machine cluster are provided, and their performance parameters are listed in **Table 2**.

Table 2. The hardware configuration of the physical hosts.

Host Number	CPU (MIPS)	REM (GB)	Hard Disk (GB)	Bandwidth (MB/s)
Physical host 1	2000	2.0	200	1000
Physical host 2	1500 (dual core)	8.0	300	1000
Physical host 3	1000 (dual core)	2.0	200	1000
Physical host 4	2000 (dual core)	8.0	500	2000
Physical host 5	2000	4.0	300	1000

After executing the virtual machine cluster deployment method based on graph cuts theory described in this article, the virtual machine cluster is segmented to form 2 sub-clusters, which are deployed to physical host 2 and physical host 5, as schematically shown in **Fig. 6**.

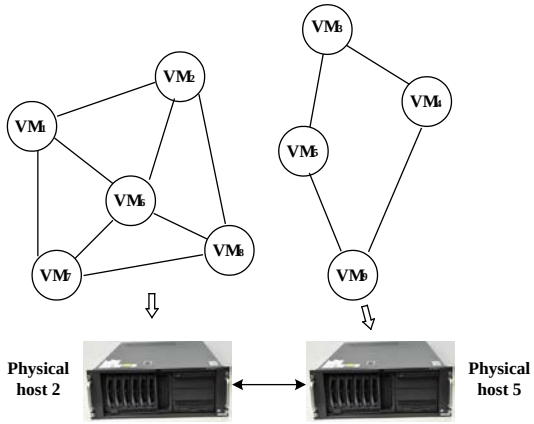

Fig. 6. The segmentation of a virtual machine cluster

In the 2 virtual machine clusters segmented based on graph cuts theory, the first cluster includes virtual machine 1, virtual machine 2, virtual machine 6, virtual machine 7, and virtual machine 8; the second cluster includes virtual machine 3, virtual machine 4, virtual machine 5, and virtual machine 9. The total CPU resource requirement, memory resource requirement, hard disk resource requirement, and bandwidth requirement of the first cluster are 2200 MIPS, 4.6 GB, 70 GB, and 280 MB/s, respectively, while those of the second cluster are 1400 MIPS, 2.4 GB, 45 GB, and 240 MB/s, respectively. The physical hosts that are able to accommodate the requirements of the deployment of the first cluster are physical host 2 and physical host 4, while those for the second cluster are physical host 2, physical host 4, and physical host 5. If physical host 4 configures 2 clusters, then there are still numerous idle resources. Thus, the first virtual machine cluster is deployed on physical host 2, and the second virtual machine cluster is deployed on physical host 5.

After the original cluster is segmented, the communication between virtual machine 2 and virtual machine 6 and the communication between virtual machine 8 and virtual machine 9 are realised through the communication between physical host 2 and physical host 5.

To further examine the difference in efficiency between the cluster-based deployment strategy after the virtual machines have clustered and the single virtual machine deployment, the resource constraint-based single virtual machine deployment method and the load balancing-based single virtual machine deployment method are selected as comparison algorithms for investigating the change in bandwidth requirements after the virtual machines

are deployed on the physical hosts. The two methods are selected for this experiments, more details please refer to the reference [28]. After the virtual machine deployments are executed following the 3 methods, the bandwidth requirement changes are shown in **Fig. 7**.

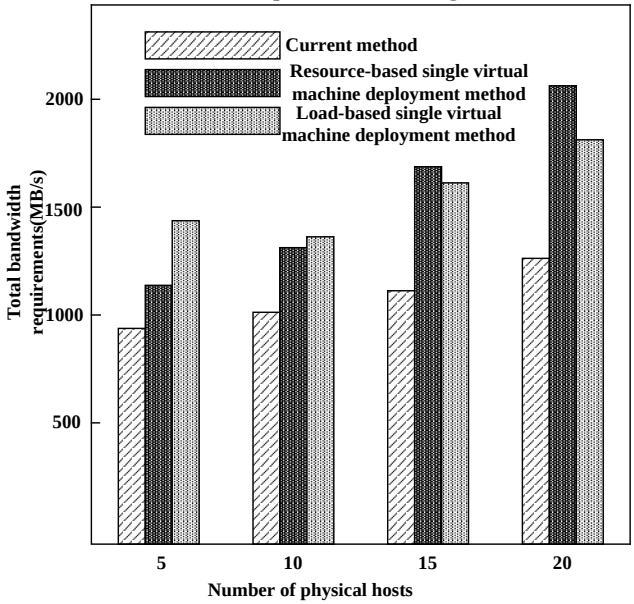

Fig. 7. Comparison of total bandwidth requirements of different methods

As shown in Figure 7, with the increasing number of physical hosts required by the virtual machine deployments, the total system bandwidth requirements under the 3 methods all increase. In relation to the 2 methods of single virtual machine deployments, the graph cuts theory-based virtual machine cluster deployment method described in this study has not only a smaller absolute value of the total bandwidth requirement but also the slowest increasing trend of bandwidth requirements with the increasing number of physical hosts.

Below, a further comparison of the average resource utilization on the physical host after three methods to complete deployment, the experimental results are shown in **Fig. 8**.

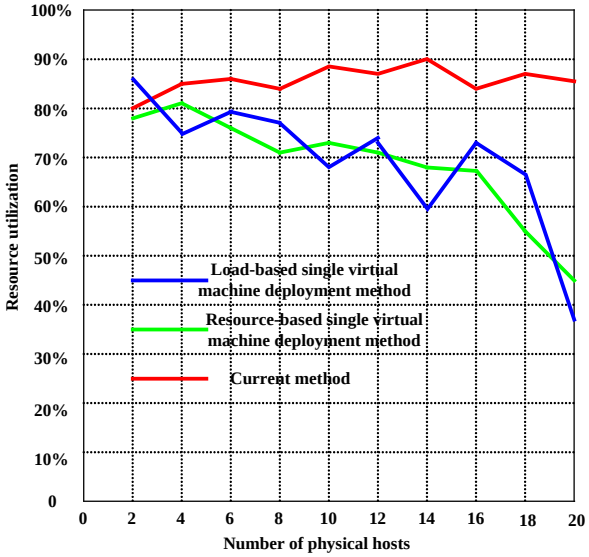

Fig. 8. Experimental results of utilization rate of resources

We can be seen from the figure 8, in the initial stage of virtual machine clusters deployment, While physical host that participating in configuration quantity is small, physical host resources utilization rate is very high after three methods to complete deployment. Along with the advancement of the deployment process, physical host resources utilization rate has remained at about 85% in our method, The single virtual machine deployment method based on resource constraints and the single virtual machine deployment method based on load balancing are present a sharp drop trend. After the complete deployment, resource utilization of the two single virtual machine deployment methods are below 50%.

The dynamics of clusters demand periodically updated clustering, and the clustering process and segmentation process of the virtual machine clusters are asynchronous, which means that the competition for resources during virtual machine deployment is reduced. Moreover, when the platform in the cloud data centre deploys virtual machine clusters, the query space in the deployment is reduced, leading to a significant reduction in deployment time, which leads to better scalability; additionally, the load balancing and service quality of the cloud data centre are enhanced.

6. Conclusion

In this study, the feasibility and related attributes of dividing virtual machine resources in a cloud platform, using the cluster as the management unit, are investigated. First, the definitions of virtual machine clusters are compared, and our version is presented; then, the feasibility of dividing using the virtual machine cluster as the management unit is discussed, the isomorphism and reconfigurability of the cluster are proven, and finally, the dynamics of the cluster are further investigated and experimentally verified. Because the virtual machine clusters exhibit no competition for resources among each other and are not dependent on the performances of application systems/components, once a virtual machine cluster has learned the optimal configuration strategy, it can readily migrate the learned knowledge to other similar virtual machine clusters, which is an effective way to simplify complex learning problems. Therefore, considering virtual machine management using the virtual machine cluster as the basic management unit is not only feasible but also applicable to the optimization management of virtual resources in a large-scale cloud computing environment. In the future, we plan to use the virtual machine cluster as the basic management unit to consider the optimization configuration of the parameters of the application systems/components of virtual machine resources. This approach takes into consideration both the complex relationship of competition for resources among different virtual machines on the same physical host and the complex relationship of performance dependencies among different components of application systems with a multi-layer structure, which straightens out the interwoven relationships among the virtual machines, among the application systems/components, and among the virtual machines and application systems/components in a large-scale cloud computing environment.

References

[1] R. Buyya, C S. Yeo, S .Venugopal, J. Broberg and I. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, no. 6, pp. 599-616, 2009. Article (CrossRef Link)

[2] M. Stillwell, D. Schanzenbach, F. Vivien, and H. Casanova, "Resource allocation algorithms for virtualized service hosting platforms," *Journal of Parallel and Distributed Computing*, vol. 70, no. 9, pp. 962-974, 2010. Article (CrossRef Link).

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, and A. Konwinski, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50-58, 2010. Article (CrossRef Link).

[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*, vol. 22, no. 6, pp. 1045-9219, 2011. Article (CrossRef Link).

[5] S. Chaisiri, B S. Lee and D. Niyato, "Optimization of resource provisioning cost in cloud computing," *IEEE Transactions on Services Computing*, vol. 5, no. 2, pp. 164-177, 2012. Article (CrossRef Link).

[6] Kourai, Kenichi., Chiba and Shigeru, "Fast software rejuvenation of virtual machine monitors, " *IEEE Transactions on Dependable and Secure Computing*, vol. 8, no. 6, pp. 839-851, 2011. Article (CrossRef Link).

[7] Y. W. Ahn, A. M. K. Cheng, J. Baek, M. Jo and H. H. Chen, "An Auto-Scaling Mechanism for Virtual Resources to Support Mobile, Pervasive, Real-Time Healthcare Applications in Cloud Computing," *IEEE Network*, vol. 27, no.5, pp. 62-68, 2013. Article (CrossRef Link).

[8] I. Foster, T. Freeman, K. Keahy, D. Scheftner, B. Sotomayer and X. Zhang, "Virtual clusters for grid communities," *in Proc. of 6th IEEE International Symposium on Cluster Computing and the Grid*, pp. 513-520, 2006. Article (CrossRef Link).

[9] F.Doelitzscher, M. Held, C. Reich and A. Sulistio, "Viteraas: Virtual cluster as a service," in *Proc. of 3th IEEE International Conference on Cloud Computing Technology and Science*, pp. 652-657, 2011. Article (CrossRef Link).

[10] M C. Golumbic, "Algorithmic graph theory and perfect graphs," *Elsevier*, Holland, 2004.

[11] J A. Bondy, U S R. Murty, "Graph theory with applications," *Macmillan*, London, 1976.

[12] M. Murphy, B. Kagey, M. Fenn and S. Goasguen, "Dynamic provisioning of virtual organization clusters," in *Proc. of 9th IEEE/ACM International Symposium on Cluster Computing and the Grid*. pp. 364-371, 2009. Article (CrossRef Link).

[13] M A. Murphy, S. Goasguen, "Virtual Organization Clusters: Self-provisioned clouds on the grid," *Future Generation Computer Systems*, vol. 26, no. 8, pp. 1271-1281, 2011. Article (CrossRef Link).

[14] Amazon elastic compute cloud (amazon ec2). Http://aws.amazon.com/ec2/

[15] WANG Guangbo, MA Zitang, SUN Lei, "Deployment of virtual machines with clustering method based on frame load awareness," *Journal of Computer Applications*, vol. 33, no. 5, pp. 1271-1275, 1288, 2013. Article (CrossRef Link).

[16] Y. Zheng, "Research on Virtual Machine cluster Deployment Algorithm in Cloud Computing Platform," *Hunan normal university*.

[17] X. Wei, H. Li, K. Yang and L. Zou, "Topology-aware Partial Virtual Cluster Mapping Algorithm on Shared Distributed Infrastructures," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 10, pp. 2721-2730, 2014. Article (CrossRef Link).

[18] Y. Huang, R. Yang, L. Cui, T. Wo, C. Hu and B. Li, "VMCSnap: Taking Snapshots of Virtual Machine Cluster with Memory Deduplication," in *Proc. of 8th IEEE International Symposium on Service Oriented System Engineering*, pp. 314-319, 2014. Article (CrossRef Link).

[19] X. Zhao, J. Yin, Z. Chen, X. Lu, "Distance-aware virtual cluster performance optimization: A hadoop case study," in *Proc. of 2013 IEEE International Conference on Cluster Computing (CLUSTER)*, pp. 1-8, 2013. Article (CrossRef Link).

[20] W. Zhou, S. Yang, J. Fang, , X. Niu and H. Song, "Vmctune: A load balancing scheme for virtual machine cluster using dynamic resource allocation," in *Proc. of 9th International Conference on Grid and Cooperative Computing*, pp. 81-86, 2010. Article (CrossRef Link).
[21] Cui. Lei, Li. Bo, Zhang. Yangyang, Li. Jianxin, "HotSnap: A Hot Distributed Snapshot System for Virtual Machine Cluster," in *Proc. of 27th Large Installation System Administration Conference*, pp. 59-74, 2013.
[22] C. T. Yang, J. C. Liu, K. L. Huang and F. C. Jiang, "A method for managing green power of a virtual machine cluster in cloud," *Future Generation Computer Systems*, vol. 37, no. 7, pp. 26-36, 2014. Article (CrossRef Link).
[23] C Y. Tseng, K Y. Liu and L T. Lee, "Enhance the Performance of Virtual Machines by Using Cluster Computing Architecture," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 11, no. 5, pp. 2553-2558, 2013. Article (CrossRef Link).
[24] Cui. Lei, Li. Jianxin, Wo, Tianyu, Li, Bo, Yang. Renyu, Cao. Yingjie, and Huai.Jinpeng, "HotRestore: a fast restore system for virtual machine cluster," in *Proc. of 28th USENIX conference on Large Installation System Administration*. pp. 1-16, 2014.
[25] Juve G and Deelman E, "Wrangler: virtual cluster provisioning for the cloud," in *Proc. of 20th international symposium on High performance distributed computing*. pp. 277-278, 2011. Article (CrossRef Link).
[26] Boykov Y, Veksler O and Zabih R, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222-1239, 2001. Article (CrossRef Link).
[27] Chen W K, "Applied graph theory," *Elsevier*, Holland, 2012.
[28] R. N. Calheiros, R. Ranjan, A. Beloglazov, Rose. De, A. F. César and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: practice and experience*, vol. 41, no. 1, pp. 23-50, 2011. Article (CrossRef Link).

Zhiping PENG received his B.S. and M.S. degrees from China University of Petroleum (HuaDong) and Huazhong University of Science and Technology in 1996 and 2001, respectively, and received the Ph.D. degree in Computer Application from South China University of Technology in 2007. He is currently a professor at Guangdong University of Petrochemical Technology. He has published more than 50 papers in refereed journals and conference proceedings. His research interests include cloud computing, machine learning and multi-agent system.

Bo XU received the M.S. degrees in Computer Science and Technology from Hunan University in 2009; He is currently a lecturer at Guangdong University of Petrochemical Technology and he is a Ph.D. candidate at South China University of Technology. He is a Senior Member of China Computer Federation (CCF), member of Computer Applications Professional Committee of CCF, IEEE member and ACM member. He has published more than 30 papers in refereed journals and conference proceedings. His research interests include cloud computing, computational intelligence, and multi-agent system.

Antonio Marcel Gates received his B.S. and M.S. degrees from Hawaii Pacific University in 2005 and 2010, respectively, and now he is a Ph.D. candidate at Hawaii Pacific University. He has published more than 10 papers in refereed journals and conference proceedings. His research interests include cloud computing and computational intelligence.

Delong Cui received his M.S. degree in communication and information system in Southwest Jiao tong University in 2008. He is currently an associate professor at Guangdong University of Petrochemical Technology. He has published more than 20 papers in refereed journals and conference proceedings. His research interests include cloud computing and reinforcement learning.

Weiwei LIN received his B.S. and M.S. degrees from Nanchang University, in 2001 and 2004, and received the Ph.D. degree in Computer Application from South China University of Technology in 2007. He is currently an associate professor at the same university. He has published more than 40 papers in refereed journals and conference proceedings. His research interests include distributed computing, cloud computing, and big data.