A Virtualized Bandwidth Resource Allocation Scheme To Improve the Transmission Performance in Cloud Computing

Shin-Jer Yang
Dept. of Computer Science & Information Management
Soochow University, Taipei, Taiwan
e-mail:sjyang@csim.scu.edu.tw;

Yi-Hsuan Chen
Dept. of Computer Science & Information Management
Soochow University, Taipei, Taiwan
e-mail:00356009@scu.edu.tw;

***Abstract*—This paper is mainly studied on one of key technologies of cloud computing: virtualization. From previous researches, most of the virtualized environment configurations are focused on the host hard disk space management, memory configuration, CPU usage, but seldom studied in terms of network bandwidth, therefore there isn't a good mechanism can provide efficient bandwidth allocation. In the absence of efficient control in bandwidth allocation, often because the number of users of the system began to grow and causes the virtual machine not offer the bandwidth, that will make the user longer wait time, and also degrade the Quality of Service of Cloud platform or system. Based on NETSHARE method, we enhance NETSHARE to propose a new virtualized bandwidth resources allocation scheme called VBRAS in this paper. The purposes of VBRAS are to dynamically allocate server's bandwidth for users and take into account the large amount number of users as well as network speed on the Client and Server. Also, the experimental results indicate that the VBRAS can obtain 20% lower convergence frequency and 8% higher throughput than the NETSHARE. Consequently, the VBRAS will offer real-time adjustment and fair allocation of bandwidth resource to improve the transmission performance under a large amount of users into the cloud platform.**

Keywords- Cloud Computing, Virtualization, Virtualized Bandwidth Resource Allocation Scheme, Bandwidth

I. INTRODUCTION

Cloud computing is optimization of computing resources, storage, network, hardware, software and related platform resources via virtualization as well as measurable and chargeable service type, is kind of service platform for users to access anytime via network distribution[5]. However, the point of cloud computing is rapid and flexible separation including network, computing, storage space related resources, to allocate required resources to users immediately. To that extent, what are special requirements for the network in cloud computing comparing to the network of current environment?

As far as network traffic variation that is hard to predict, a need of flexible allocation of resources in cloud environment has made network type of entire cloud data center become more dynamic and is difficult to predict. Owing to fixed number of staff and proficiency in operation type of traditional self-established data center, it will be easier to manage and predict network traffic of entire data center; but in a cloud environment, it is hard for service providers to evaluate operation type of respective users, which shall be adjusted and changed as per resource consumption condition, enabling bigger variation to entire cloud network. Also, most of measurable resources on current virtual machines lay emphasis on resource allocation of memory, CPU and Storage and lack of a discussion on network resource. Therefore, allocation and management of bandwidth resource is one the essential issues of cloud computing [2].

Based on the NETSHARE method [9], we enhance NETSHARE features to propose a new scheme, called VBRAS (Virtualized Bandwidth Resource Allocation Scheme). To observe changes in data volumes of various virtual machines via fixed interval *t* second, to real-time adjust limitation of bandwidth, to have bandwidth resources used in a more effective way. However, there is no discussion on the effect enabled by the number of users, as a consequence, the primary purposes of this paper are as follows:

(1) In consideration of no definite allocation strategy of bandwidth resources, it can take moderate adjustment on bandwidth via analysis of Internet usage status on the current various virtual machines for improving poor allocation of bandwidth resources.

(2) We propose an efficient allocation method for virtualized bandwidth resources regarding network resources that were seldom discussed in virtual environment.

(3) Based on NETSHARE method, we add number of users and data volume as 2 factors to propose the VBRAS for improving the transmission performance.

(4) The VBRAS can find out a proper coefficient as ceiling for bandwidth adjustment upon growth percentage of users.

(5) We use VBRAS scheme to perform simulation experiments and compare with NETSHARE to prove that effective use of bandwidth resources can be enabled through real-time and dynamic adjustment on bandwidth limit when it exists bigger variation in number of users.

(6) Via VBRAS scheme, bandwidth resource required by every virtual machine could be adjusted to steady mode within shorter time than the NETSHARE method.

The rest of this paper is organized as follows. Section II surveys and discusses the relevant literature required for this paper. Section III comprises the VBRAS scheme and designs its pseudo-code. Section IV lists the simulation environments and procedures for experiments. Section V examines the comparative results and makes the analysis between VBRAS and NETSHARE. Section VI draws the conclusion and indicates the future direction.

© 2013. The authors - Published by Atlantis Press

II. RELATED WORKS

A. *Cloud Computing*

According to the NIST definition of Cloud Computing, there are five characteristics including On-demand self-service, Broad network access, Resource pooling, Rapid elasticity and Measured service. Also, Cloud computing platform covers 3 service models and 4 deployment types, which are shown as Figure I.

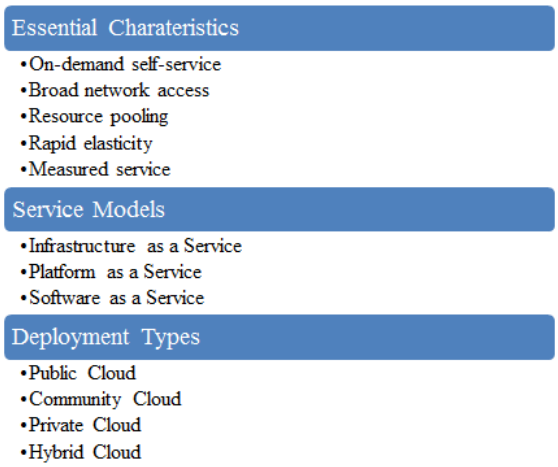

Figure I. Cloud Computing Architecture

These Cloud services are broadly divided into 3 models, the primary features of these 3 models are listed:

(1) IaaS (Infrastructure as a Service): Iaas is aggregate of hardware resources and related management functions after virtualization, to realize automation of internal process and optimization of resource management after having computing, storage and internet related resources abstracted, and to further provide dynamic and nimble infrastructure service.

(2) PaaS (Platform as a Service): Provide an environment of development, operation, management and monitoring for cloud applications, which could be claimed as optimized "cloud middleware", in which superior platform layer design could satisfy cloud's requirements in terms of expandability, usability and safety aspects.

(3) SaaS (Software as a Service): These applications are built on resources provided by infrastructure layer and environment provided by platform layer and to be rendered to client via internet. Applications provided by this layer could allow users to access service via multiple internet connection devices, with browser or internet connection interface opened simply without worrying installation and upgrade of software. As far as application developer is concerned, they could conveniently conduct software deployment and upgrade with no need to manage or control cloud structure of bottom layer such as internet, server, operating system and storage etc.

In addition, 4 main deployment types of cloud services are listed as follows.

(1) Public Cloud: The public cloud can sell services to anyone on the Internet. Currently, Amazon Web Services is the largest public cloud provider. The public cloud environment is provided by one cloud provider, and more than one organization can access it

(2) Community Cloud: The Infrastructure of community cloud can be owned by multiple organizations and be supported by common issues. This cloud type should consider security needs. The cloud environment is established by several organizations which have similar requirements and seek to share the infrastructure

(3) Private Cloud: The private cloud is a proprietary network or a data center that supplies hosted services to a limited number of people. The private cloud environment is provided or constructed by one company, and this environment is used for only one organization

(4) Hybrid Cloud: The hybrid cloud environment is integrated with two or more cloud models. The use of physical hardware and virtualized server instances together to provide a single common service. When a service provider uses public cloud resources to create their private cloud. But, hybrid cloud is easy to migrate data and application via standardization and dedicated technique

However, private or public, the goal of Cloud Computing is to provide easy, scalable access to computing resources and IT services.

B. *Virtual Machine*

Virtualization is to quantize server resources such as network, memory, CPU or storage space after abstraction and conversion, user accounts will be able to apply these resources upon a manner better than original configuration. The new virtualization part of these resources is subject to no limitations on installation manner of existing resources, geography or physical configuration.

SaaS is to turn application software into quantified resources and allocate accordingly, i.e. "Software as a Service", and to charge according to the volume used, this volume could be calculation ability, or function, or use time. Via a concept of user charge, for users, on the one hand, could use the least money to complete one mission, on the other hand, fulfill a spirit of no resource waste to achieve energy saving and carbon reduction effects. But in addition to software, what is the level down below? If you would like to establish a website by yourself, you could choose virtual server, independent server, server co-location or a stack of servers. In which a server may have multiple CPUs and 16GB RAM, from little CPU and RAM on virtual server occupied to Core 2 Quad CPU and 16GB RAM on independent server, if spec. demand is bigger than virtual server but smaller than independent server, how do we choose it from? Processing speed would be not enough if virtual server is chosen while it will be a waste if the independent server is selected. Therefore, Virtual Private Server (VPS) has then appeared. The so-called virtual

independent server is a virtual machine (VM) which occupies part of entire server's resources and provides a complete environment for management.

C. NETSHARE

Vinh the Lam and others execute 3 NETSHARE mechanisms. First, dependence on TCP and fair queuing, just alter allocation and make response to changes caused by few round-trip delay. We show idea of how this could extend to more application software and use NETSHARE randomly. Second mechanism makes up the first mechanism to process UDP. Lastly, the third mechanism uses centralized allocation to provide a more general bandwidth allocation.

(1) Stochastic NETSHARE

Since the switches may support limited DRR queues, NETSHARE scales to a large number of application classes by stochastic weighted max-min fair sharing, which is a generalization of McKenney's Stochastic Fair Queuing.

Since queue sharing allows a small weight service to steal more bandwidth, authors propose a mixture of random and weight-based allocation as follows. First, grouping services based on weight classes (say all services of weight 1, all of weight 2, all of weight 4, etc.). Then map each weight class into a set of queues and randomly assign services to a specific queue within each set[4].

(2) Rate Throttling for UDP

Each host is instrumented with a rate throttling shim layer just below UDP. As illustrated in Figure II, suppose H1 sends traffic at 10 Gbps to another host H4. The shim layer at H4 measures received traffic of 1 Gbps from H1. This is sent back to the corresponding rate throttling layer at H1 which rate-limits the traffic at close to 1 Gbps. Furthermore, to allow legitimate rate growth (e.g., H1 could grow to 2 Gbps if H2 disappears), set the throttled rate to somewhat more than the measured rate to allow ramp-up. Also, do not let the rate to fall below a threshold to avoid long ramp-up of small flows.

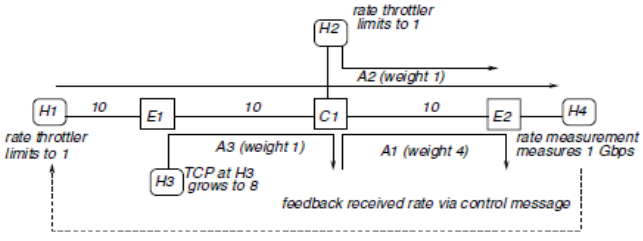

Figure II. Simple Fair Queuing.

(3) Centralized Bandwidth Allocator

In this sub-section, we describe a centralized bandwidth allocator to allow advanced bandwidth allocation policies beyond max-min fairness and there are four steps: Step 1. Rate Measurement: The rate of each flow for each service is measured at either the switches or at the hosts in intervals of *T* seconds and used to predict a demand for the next interval. Step 2. Rate Reporting: The predicted rates are sent to a centralized bandwidth allocator that is also supplied with the service weights and the topology via routing updates. Step 3. Centralized Calculation: The centralized allocator calculates rates for each flow and each service and sends back rate updates to the switches or hosts. Step 4. Rate Enforcement: Token bucket rate-limiters are used at the hosts or ingress switch ports to limit the rates to the calculated rates.

III. Operating Procedure and Design Issues of VBRAS

A. Operating Procedure

In this sub-section, we validate and enhance NETSHARE method and then propose a new bandwidth adjustment scheme, called VBRAS, as to dynamic bandwidth adjustment of increasing or decreasing in number of users on Cloud servers of VMs. The VBRAS algorithm can automatically adjust bandwidth used by users based on network load of current VM and determine whether adjustment on VM is allowed by considering increasing or decreasing in number of users. Hence, there are following 3 design concepts for VBRAS algorithm.

(1) Dynamic bandwidth adjustment: The system can adjust bandwidth ceiling according to current network traffic load status on the cloud server.

(2) Bandwidth adjustment with changing in number of users: The VBRAS scheme will conduct adjustment on the VM based on changes in number of users.

(3) When data volume of VM and Cloud users are all ste to zero, the idle mode of Cloud server will be determined without conducting any adjustment.

The operation process of VBRAS scheme is shown in Figure III, with data volume of Client and Server and amount of current users recorded firstly, the system will use fixed interval *t* second to record above 3 statistics and observe changes in statistics every time, to determine coefficient for bandwidth adjustment and adjust ceiling of bandwidth moderately. If data volume of Server and Client are all set to zero, then coefficient of bandwidth will be adjusted to 0, while bandwidth limit will no more be adjusted, and to be re-adjusted until any a variation in data volume are observed.

B. Design of VBRAS Algorithm

Based on the VBRAS operations as depicted in Figure III, we design the pseudo codes of VBRAS algorithm as follows.

Algorithm VBRAS()

{

Input:

int d = 10%; // The difference between the amount of data

int p; // The difference between the number of users

int γ_I; // increase the bandwidth factor

int γ_D; // reduce the bandwidth factor

int L; // measuring the amount of server's data at T_{n-1}

int C; // measuring the amount of server's data at T_n

int R_P; // change rate of the number of users

int R_N; // current network speed
int P_L; // measuring the number of client's users at T_{n-1}
int P_C; // measuring the number of client's users at T_n
int S = 0; // the times of adjusting to reach steady state
boolean A = True; // determine whether to allow the adjustment of the bandwidth
Output:
To adapt γ_I, γ_D to control bandwidth limit
Method:
If (C==L=='0') then γ_I=γ_D=0
Else
If (| (L - C) / L | ≧ d) then
    If (C –L > 0) then
        If ((P_C - P_L) / L > R_P) then
            A ← True
            R ← C * (1 +γ_I)
            S++
        Else
    If ((P_C - P_L) / L > R_P and P_C - P_L > 0) then
        A ← True
        R ← L * (1 + γ_D)
        S++
    Else
        A ← True
        R ← C * (1 + γ_D)
        S++
    L ← C
}
END VBRAS.

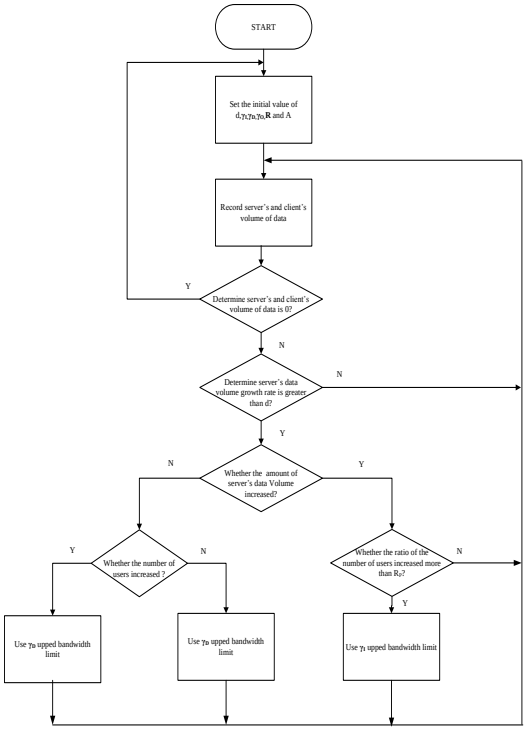

Figure III. Operating Procedure of VBRAS

IV. Simulations Environment and Procedure Design

A. Simulations Environment

The experiment of this paper will require two systems, which can be regarded as Server and Client, respectively. The Server will install VMWARE 5.0, with 3 VMs installed internally, in which 3 VMs will have their respective Service Level Agreement (SLA), while the Client is Windows 7 based system, to connect to VMWARE 5.0 environment of Server via the VMWARE VSphere Client installed on its computer. Apply NETSHARE algorithm and VBRAS algorithm proposed by this paper on the Server respectively to analyze different data volumes caused by different number of users, and to allocate bandwidth based on SLA of each VM to be compared the transmission performance by NETSHARE and VBRAS algorithms.

B. Simulation Procedures

In this paper, we will analyze and observe the simulation results by following 2 KPIs.

(1) Convergence frequency：the so-called Convergence frequency refers to number of adjustment from bandwidth ceiling adjustment of VM to a state requiring no adjustment of bandwidth ceiling, the fewer the Convergence frequency, the better the adjustment efficiency.

(2) Throughput：TX and RX data volume of network speed within certain period of time in order to understand changes in network traffic.

Also, we design the simulation procedures as following steps for conducting experiments to be compared with VBRAS and NETSHARE algorithms.

- Step1. First, the VBRAS will record VM data traffic and the number of users every t seconds. If the data volume has reached the upper limit of default bandwidth, it will be able to adjust the bandwidth limitations γ_I
- Step 2. Every t seconds, the VBRAS will once again record the data volume of each VM, and calculate the difference value: d of data volume for each VM between T_n and T_{n-1}. If d was greater than the default value, then VBRAS can adjust the bandwidth limitations according to the number of users and the growth rate of data volume.
- Step 3. If d is zero, the VBRAS determined that VM is into a stable state, in addition, if the data volume of the VM was zero, then VBRAS will temporarily not adjust bandwidth limitations, but it will still continue to record the data volume and the number of users of each VM.

V. Results Analysis

First, we perform a simulation to collect the data of the current data volume and the number of users on each VM once every 20 seconds. Then, NETSHARE and VBRAS will adjust bandwidth limitation dependent on current data volume and number of users once every 100 seconds as shown in Table I. The simulation result from Table I

indicates that the VBRAS can obtain 20% lower convergence frequency than the NETSHARE.

Second, we also perform a simulation to measure the data volume of each VM. Figure IV shows the data volume through the beginning to the 300th second and then reach the upper of the bandwidth limitation, so they have raised about 10% of their bandwidth limitation. At the 400th second, because the growth rate in number of users does not exceed the value of R_p, so VBRAS algorithm is to reduce the bandwidth limitation based on the data log at 200th second. Thereafter, the measured data volume cannot reach the bandwidth limitation, we call that as a steady state under this situation. Hence, we did not have to adjust bandwidth limitation on this situation. From the beginning of the experiment to reach a steady state, NETSHARE algorithm spends five times adjustments to reach steady state, and VBRAS only spends four times to reach steady state. The simulation result from Figure IV indicates that the VBRAS can have 8% higher throughput than the NETSHARE.

Table I. Total adjusting times with NETSHARE and VBRAS

Time(second)	0s	100s	200s	300s	400s	500s	600s
Number of users	0	50	100	200	210	450	500
Bandwidth limit of NETSHARE	50KB	55KB	60KB	66KB	60KB	55KB	55KB
Total adjusting Times	0	1	2	3	4	5	5
Bandwidth limit of VBRAS	50KB	55KB	60KB	66KB	55KB	55KB	55KB
Total adjusting times	0	1	2	3	4	4	4

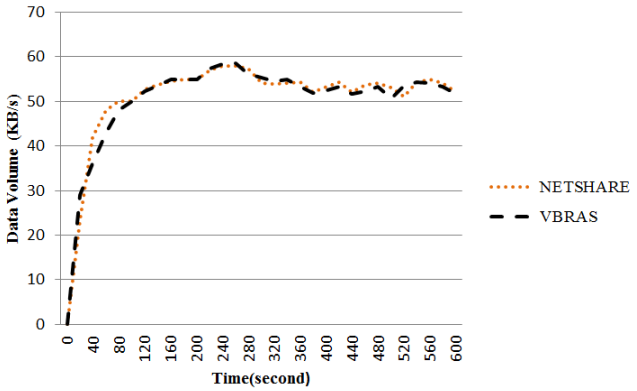

Figure IV. Data volume of NETSHARE and VBRAS

VI. CONCLUSION

The paper has analyzed virtualization problems that cloud environment encounters and targeted to design a new dynamic adjustment on virtualized bandwidth resources, with considerations of number of users and data volume to propose a new bandwidth allocation scheme: VBRAS. Use of this scheme will help improve different demands on bandwidth resources out of data volume caused by changes in number of users, enabling inefficient allocation of bandwidth resources to those VMs that are truly in need. Hence, we can real-time adjust bandwidth limit via VBRAS algorithm to allow bandwidth resources to be more efficient and more effective.

The main contribution of this paper is to provide a new scheme for virtualized network resources management. Based on the simulations, the final results indicate that the VBRAS can obtain 20% lower convergence frequency and 8% higher throughput than the NETSHARE. Use of proposed VBRAS can help allocate bandwidth to users more efficient when system faces large level of changes in number of users and data volumes, to allow users to use the system smoothly and to further avoid abnormal and unsmooth use of the system. Thus, client's loyalty to the system will then be retained. In the future, we will perform further simulations in terms of other KPIs such as average response time and wasted ratio.

ACKNOWLEDGMENT

The partial work of this paper is supported by National Science Council in Taiwan under Grant NSC 102-2410-H-031-061.

REFERENCES

[1] A. Amamou, M. Bourguiba, K. Haddadou, G. Pujolle, "A Dynamic Bandwidth Allocator for Virtual Machines in a Cloud Environment", In Proceedings of Consumer Communications and Networking Conference, pp. 99-104, 2012.

[2] Chuanxiong Guo , Guohan Lu , Helen J. Wang , Shuang Yang , Chao Kong , Peng Sun , Wenfei Wu , Yongguang Zhang, "SecondNet: a data center network virtualization architecture with bandwidth guarantees", In Proceedings of the 6th International Conference, November 30-December 03, 2010.

[3] J. Shafer, "I/O Virtualization Bottlenecks in Cloud Computing Today", In Proceedings of the 2nd Conference on I/O Virtualization, pp. 5-5, 2010.

[4] Vinh The Lam, Sivasankar Radhakrishnan, Rong Pan, Amin Vahdat, George Varghese, "Netshare and Stochastic Netshare: Predictable Bandwidth Allocation for Data Centers", In Proceedings of ACM SIGCOMM Computer Communication Review, Vol 42, No. 3, pp. 5-11, 2012.

[5] Lam et al. NetShare and Stochastic NetShare: Predictable Bandwidth Allocation for Data Centers. In UCSD Tech Report 2011. http://cse.ucsd.edu/users/vtlam/netshare-TR-2011.pdf.

[6] M. Hasan Jamal, A. Qadeer, W. Mahmood, A. Waheed, J. J. Ding, "Virtual Machine Scalability on Multi-Core Processors Based Servers for Cloud Computing Workloads", In Proceedings of IEEE International Conference on Networking, Architecture and Storage, 2009.

[7] George Pallis, "Cloud Computing: The New Frontier of Internet Computing", Internet Computing Journal, Vol. 14, No. 5, pp. 70-73, Sep./Oct. 2010.

[8] Xing Pu , Ling Liu , Yiduo Mei , S. Sivathanu, Younggyun Koh , C. Pu, "Understanding Performance Interference of I/O Workload in Virtualized Cloud Environments", In Proceedings of Cloud Computing Conference, pp. 51-58, 2010.

[9] Eyal Zohar, Israel Cidon, Osnat Mokryn, "The Power of Prediction: Cloud Bandwidth and Cost Reduction", In Proceedings of the ACM SIGCOMM 2011 Conference, pp. 86-97, 2011.

[10] Amazon Elastic Compute Cloud, http://aws.amazon.com/ec2.