ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

HSAWM: An Efficient Hybrid Simulated Annealing Workload Management in Cloud Computing

V.Karpagam[1], V.Venkatesakumar[2]

M.E Student, Dept. of C.S.E., Regional Center of Anna University, Coimbatore, India[1]

Assistant Professor, Dept. of C.S.E., Regional Center of Anna University, Coimbatore, India[2]

ABSTRACT: Performance of a cloud is a critical issue and dependent on various factors that include load balancing used to distribute the load at data center.In existing researchof Hybrid Cloud Computing (HCC) determines whether workloads should reside in data center private cloud, the public cloud or a hybrid combination. It doesn't maximize utilization in the private cloud, only using the public cloud to meet unpredictable workload demand. In this paper presents hybrid simulated annealing algorithm takes maximize utilization in the private and public cloudand locating a server to the global optimum of a given function in a large dimensional search space. Meanwhile, the algorithm performs to solving the userand server allocation problems in a cloud computing environment.

KEYWORDS: Cloud computing; virtual machine; simulated annealing algorithm; load balancing.

I. INTRODUCTION

Cloud Computing, known either as online services such as Amazon AWS [1] and Google App Engine [2], or a technology portfolio behind such services, features a shared elastic computing infrastructure hosting multiple applications where IT management complexity is hidden and resource multiplexing leads to efficiency; more computing resources can be allocated on demand to an application when its current workload incurs more resource demand than it was allocated.

An intelligent workload factoring service is designed as an enabling technology of the hybrid cloud computing model. Its basic function is to split the workload into two parts upon (unpredictable) load spikes, and assures that the base load part remains within plan in volume, and the trespassing load part incurs minimal cache/replication demand on the application data required from it. This simplifies the systemarchitecture for the trespassing load zone and significantly increases the server performance within it. As for the base load zone, workload dynamics are reduced significantly; this makes possible capacity planning with low over-provisioning factor and/or efficient dynamic provisioning with reliable workload prediction.

Current data centers leverage virtualization to enhance fault and performance isolation, improve system manageability and reduce infrastructure cost. Virtualization provides an opportunity to significantly improve utilization and efficiency of the resources by clustering the workload into as few servers as possible and shutting down the rest. This leads to both significant energy savings due to powering fewer servers and to a large utilization increase on the remaining servers. However, this also dramatically increases chances of creating thermal hotspots in the system.

The rest of this paper is organized as follows. In Section 2 review the existing related work. The proposed models and descriptions are described in Section 3. Finally conclude the paper in Section 4.

II. RELATED WORK

In [3] authors measured and analyzed the workload on Yahoo! Video, the 2nd largest U.S. video sharing site, to understand its nature and the impact on online video data center design. We discovered interesting statistical properties on both static and temporal dimensions of the workload; they include file duration and popularity distributions, arrival rate dynamics and predictability, and workload stationarity and burstiness. In [4] authors explored the influence of task

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

size variability on which task assignment policy is best. Surprisingly, we find that not one of the above task assignment policies is best. In particular, we find that when the task sizes are not highly variable, the Dynamic policy is preferable. However, when task sizes show the degree of variability more characteristic of empirically measured workloads, the size-based policy is the best choice.In [5] authors discussedpossible reasons for this performance degradation. To describe a careful implementation of a multi-level direct KK-way hypergraph partitioning algorithm, this performs better than a well-known recursive-bisection-based partitioning algorithm in hypergraph partitioning with multiple constraints and fixed vertices. In [6] authors illustrated the Per-flow network traffic measurements are needed for effective network traffic management, network performance assessment, and detection of anomalous network events such as incipient DoS attacks. Explicit measurement of per-flow traffic statistics is difficult in backbone networks because tracking the possibly hundreds of thousands of flows needs correspondingly large high-speed memories. To reduce the measurement overhead, many previous papers have proposed the use of random sampling and this is also used in commercial routers (Cisco's Net Flow). The goal is to develop a new scheme that has very low memory requirements and has quick convergence to within a pre-specified accuracy. To achieve this by use of a novel approach based on sampling two-runs to estimate per-flow traffic.In [7] authors discussed a many large content publishers use multiple content distribution networks to deliver their content, and many commercial systems have become available to help a broader set of content publishers to benefit from using multiple distribution networks, which we refer to as content multihoming. In this paper, we conduct the first systematic study on optimizing content multi-homing, by introducing novel algorithms to optimize both performance and cost for content multi-homing. In particular, we design a novel, efficient algorithm to compute assignments of content objects to content distribution networks for content publishers, considering both cost and performance. We also design a novel, lightweight client adaptation algorithm executing at individual content viewers to achieve scalable, fine-grained, fast online adaptation to optimize the quality of experience (QoE) for individual viewers.

III. PROPOSED ALGORITHM

The proposed methodology designed to reduce the workload of the server in banking sector using max-min algorithm. User request or task is dividing into number of subtask and each sub-task executed by different server depends upon response time of the server.

A. *Cloud Resource Manager and User Allocation*

The cloud resource manager of the centralized cloud managementsystem stores the global service task load information collected from server groups, and decides the quantity of client's requirements assigned to each server cluster so that the load of each server cluster is distributed as balanced as possible in terms of the cost of transmitting message data between server clusters and clients. The decision of assignment is based upon the characteristics of different service requests and the information collected from server clusters. The user allocation category has two types. Administrator and user play an important position. Administrator has the responsibility of creating an account to the new user and maintaining the account information of all the existing users. To create a new account the user has to give their information to the administrator. If it is valid information then the account will be created by the administrator to continue the transaction. If an account is created successfully, the existing user can view their own account information and they can transfer amount to another account.

B. *Response Time Calculation*

In response time calculation a user can requests the data to the server. If the server has the information related to the user request, it will respond the user. The response time of each task is calculated by $R_T=R_{ST}-R_{RT}$. Here RT denotes response time of the requested task, R_{ST}denotes request starting time and R_{RT} denotes response receiving time.

The response time of each and every server may vary from one server to another server. Here the response time of the entire available server is calculated to find the minimum response time.The primary allocated task in a cloud virtual machine is not an opening job, data from parent tasks have to be moved to the virtual machine before the assignment can run, and thus, it needs to be provisioned before the primary time of its first task.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

There are three service time policies:
1. Nearest Data center Policy
2. Normalize reply instance strategy
3. Dynamically variable path work load.

Nearest Data center strategy: A data center which contains the smallest amount of nearness from the client is preferred. Proximity is period of small system latency. The additional data centers containing similar closeness then it will choose data center at random to stability the load.

Normalize reply instance strategy: This strategy recognizes the nearest data center with existing rule when Nearest Data center's presentation begins corrupting it approximations present reply time for all data center then finds which containing low predictable response time. Although may be half of chances for the collection of nearest and best datacenter.

Dynamically variable path work load: The method is an addition of nearest Data center strategy where the steering reason is analogous. But it has more dependability dimension of application deployment focused on the load. It moreover enhancing otherwise reduces a number of virtual machines consequently. This will be done pleasing below reflection the present dispensation times and finest dispensation time ever realized.

C. *Load balancing Information Repository*

An information repository is an easy way to organize a secondary tier of data storage. It is used to reduce the maintenance workload. The response time and bandwidth information of all the servers are stored in this information server. The response time of previous task history is maintained in the information repository. The response time of new task is taken according to the previous task. If the server has minimum response time then it will execute the task quickly to compare with other servers.The load balancing Throttled algorithm is entirely focused on virtual machine. The user's primary appealing the load balancer to make sure accurate virtual machine which contact that load simply and complete a procedure which is given by the user. The corresponding model, Customer stats the requests load balancer to search an appropriate Virtual Machine is to execute necessary operation. The entire completing time is predictable in 3 phases. In the first one the message of the virtual machines and they will be unused waiting for the scheduler to allocate the jobs in the row, once jobs are allocated, the virtual machines in the cloud will begin dispensation, which is the next phase, and lastly in the clearout or the destruction of the virtual machines. The throughput of calculating model can be predictable as the full number of jobs are completed within a time distance without consider the virtual machine configuration time and destruction time.

D. *Cloud Data Server Allocation*

The cloud data server will get the bandwidth and minimum response time form the information server. The information taken from the information server is sorted in a particular order to find the minimum response time server. If the server has minimum response time then it will be assigned to execute the first task. The next minimum response time server will be assigned to execute the next task and so on. If the tasks are executed successfully then a mail alert will be sent to the user to grant the successful transaction. The proposed simulated annealing algorithm, an initial solution is always selected from the variation range of each of the parameters at random. In this work method the initial solution s_0 can be as follows: first maintain a list of all the VMs in descending order according to their loads, allocating them to the physical servers in sequence. Then, map the first virtual machine in the list which has the most workload to a physical machine which has the most residual capacity, and repeat this process for the next virtual machine until all the VMs have been allocated to a physical server. At last, obtain an initial solution s0 as an input of the simulated annealing load balancing algorithm.

IV. PSEUDO CODE

Step 1: Input: cloud user message, message Length *l*
Step 2: Initialize;

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Step 3:M:=0;
repeat
Process (config.load→config.load.idx, ΔVM_{ij});
ifΔVM_{ij}<=0 then accept else
ifexp(-ΔVM_{ij} /l) >random [0,1) then accept;
if accept then UPDATE(configuration j);
until search is approached sufficiently
closely;
UserM+1:=f(UserM);
M:=M+1;
until stop criterion = true (System is frozen);
end.

V. CONCLUSION AND FUTURE WORK

The proposed algorithm performs proactive workload management technology; the hybrid cloud computing model allows users to develop a new architecture where a dedicated resource platform runs for hosting base service workload, and a separate and shared resource platform serves flash crowd peak load. The proposed simulated annealing algorithm takes optimization problems like locating a server to the global optimum of a given function in a large dimensional search space.The workload Balancing algorithms are used to increase the availability of the servers, to reduce the response time of the job, to increase user satisfaction and to improve performance of the Cloud Environment. In a hybrid cloud computing model is a combination of private clouds and public clouds. By using hybrid cloud computing, the main advantage as low cost and security.

In future research, the response time of the server is calculated and maintained in the repository for the future allocation. The response time is the time difference between the task allocation time and task completion time.

REFERENCES

1. "Amazon web services," http://aws.amazon.com/.
2. "Google app engine," http://code.google.com/appengine/.
3. X. Kang, H. Zhang, G. Jiang, H. Chen, X. Meng, and K. Yoshihira, "Measurement, modeling, and analysis of Internet video sharing siteworkload: a case study," in Proc. 2008 IEEE International Conference on Web Services, pp. 278–285.
4. M. Harchol-Balter, M. E.Crovella, and C. D. Murta, "On choosing a task assignment policy for a distributed server system," pp. 31–242, 1998.
5. G. Karypis and V. Kumar, "Multilevel k-way hypergraph partitioning," in Proc. 1999 ACM/IEEE Conference on Design Automation, pp. 343–348.
6. M. S. Kodialam, T. V. Lakshman, and S. Mohanty, "Runs based traffic estimator (rate): a simple, memory efficient scheme for per-flow rate estimation," in 2004 INFOCOM.
7. H. H. Liu, Y. Wang, Y. R. Yang, H. Wang, and C. Tian, "Optimizing cost and performance for content multihoming," in Proc. 2012 ACMSIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, pp. 371–382.