Volume 5, No. 4, April 2014 (Special Issue)

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Profit Maximization via Virtual Resources Allocation in Cloud-Computing

K. Delhi Babu, D. Giridhar Kumar and M.G.Madhusudhan*

[1,2,3]Dept. of CSE

Sree Vidyanikethan Engineering College, Tirupati AP, India

kdb_babu@yahoo.com[1], giridhar.svec@gmail.com[2], madu.526@gmail.com[3]

Abstract: With increasing demand for high performance computing and data storage, distributed computing systems have attracted a lot of attention. To have a cost efficient usage of computing resources, resource scalability and on demand services are required through virtualization and distributed computing. Service charges and business costs both should known by the service provider for the profit maximization. These are determined by the multi server system configuration and the applications. Optimizations of computing and networking resources need to be jointly performed. The problem of optimal multi server configuration for profit maximization in cloud computing takes such factors as the service provider's margin and profit, the quality of service(QoS), the cost of renting, the cost of energy consumption, the workload of an application, the service level agreement(SLA), the configuration of multi server system, the amount of service. Optimization problem can be formulated and solved analytically by using M/M/m queuing model in multi server system configuration. As a framework to virtual resource mapping , a mixed integer programming(MIP) problem is formulated which relates to cost efficiency of resource mapping procedure. The link mapping can be achieved by multi commodity flow allocation problem.

Keywords: Cloud computing, SLA, multi server system, resource allocation, resource mapping, queuing model

I. INTRODUCTION

Cloud computing can be defined as the delivery of hosted services over the Internet, such that accesses to shared hardware, software, databases, information, and all resources are provided to consumers on-demand by centralized management of resources and services.

A cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and the consumers.

Cloud services include Software as a Service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (Paas). The aim of cloud computing is to allocate virtual resources that enables computing and storage data access on demand basis. For allowing more requests, cloud services has the capacity of multiplexing the physical resources among requested resources. Cloud computing and networking are the two key functionalities that are involved in the distributed clouds. Convergence between cloud and networking is more important for QoS delivery and for creation of networked cloud environments.

Cloud computing is able to provide the most cost-effective and energy-efficient way of computing resources management. Cloud computing turns information technology into ordinary commodities and utilities by using the pay-per-use pricing model [5], [6]. However, cloud computing will never be free and understanding the economics of cloud computing becomes critically important.

Three tier structure [9], a cloud computing environment consists of infrastructure vendors, service providers, and consumers. Cluster computing systems [11] including cluster nodes, cluster managers, and consumers, and Grid computing systems including resource providers, service providers, and clients are the two approaches followed. An infrastructure vendor maintains basic hardware and software facilities. A service provider rents resources from the infrastructure vendors, builds appropriate multi server systems, and provides various services to users. A consumer submits a service request to a service provider, receives the desired result from the service provider with certain service-level agreement, and pays for the service based on the amount of the service and the quality of the service.

A multi server system contains multiple servers, and such a multi server system can be devoted to serve one type of service requests and applications. An application domain is characterized by two basic features, i.e., the workload of an application environment and the expected amount of a service. The configuration of a multi server system is characterized by two basic features, i.e., the size of the multi server system and the speed of the multi server system. Service provider in cloud computing is based on two components "the income (Service charges to the users) and the cost (renting cost plus utility cost paid to infrastructure vendors)".

II. RELATED WORK

a. The distributed resource allocation problem is one of the most challenging problems in the resource management problems. The SLA based distributed resource allocation has attracted attention of the research community in the last years. Our paper considers the resource management problem in a cloud computing system. Key features of our formulation and subsequent proposed solution are that we Use a three dimensional model of the resources in the clusters, i.e., computational, storage and networking capabilities

CONFERENCE PAPER

Two day National Conference on Innovation and Advancement in Computing

Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India

Schedule: 28-29 March 2014

A mathematical formulation for the resource allocation problem in clusters is presented in [2]. The authors describe a method to find the best resource assignment in a cluster in the case that the application has certain resource requirements.

Chandra et al. [3] introduce a dynamic resource allocation method in shared clusters to minimize the overall penalty resulting from not satisfying the SLA requirements in the response time. For this optimization, online measurements of the most important parameters in the system are used to predict the next system state and to allocate resources on that basis. An economic approach to manage shared resources and minimize the energy consumption in hosting centers is described in [12].

The problem of assigning interconnected virtual nodes to the substrate network with constraints on virtual nodes and virtual links, can be reduced to the NP-hard multiway separator problem. Most of the proposed approaches decompose the problem into the node mapping phase and the link mapping phase, to reduce the overall complexity of the problem. Researchers usually employ some greedy heuristic approach for node mapping, while link mapping is performed using (k) shortest path or multi commodity flow algorithms (e.g.,[2]). Recent approaches tend to solve the two problems either simultaneously or providing some type of coordination among the two phases (e.g., [1]). In the proposed study, we follow the latter approach. The two phases are correlated in the sense that the node mapping phase facilitates the link mapping phase.

III. SYSTEM MODEL

Here, we use P[e] to denote the probability of an event e. For a random variable x, we use f_x (t) to represent the probability density function of x, and F_x (t) to represent the cumulative distribution function (cdf) of x, and x to represent the expectation of x.

A. Multiserver Model:

In cloud computing, the service requests raised by the users' are handled by the service provider by using a multiserver system. The service provider rents the multiserver system from the infrastructure vendor, who constructs and maintains multiserver system. The architecture of the multiserver system can be flexible. Examples are blade servers and blade centers where each server is a server blade [10], traditional servers where each server is an ordinary processor [11], and multicore server processors where each server is a single core. The users (i.e., customers of a service provider) sends the service requests (i.e., applications and tasks) to a service provider, and the service provider performs the requests (i.e., run the applications and perform the tasks) on a multiserver system.

Let us assume that a multiserver system S has m identical servers. Here, a multiserver system is treated as an M/M/m queuing system. Poisson stream of service requests with arrival rate , i.e., the interarrival times are independent and identically distributed (i.i.d.) exponential random variables with mean 1/ . In multiserver system S maintains a queue with infinite capacity for waiting tasks when all the m servers are busy.

The first-come-first-served (FCFS) approach is considered, then the task execution requirements (measured by the number of instructions to be executed) are i.i.d. exponential random variables r with mean $\bar{r}$. The m servers (i.e., blades/processors/cores) of S have identical execution speed s (measured by the number of instructions that can be executed in one unit of time).

Hence, the task execution times on the servers of S are i.i.d. exponential random variables x=r/s with mean $\bar{x} = \bar{r}/s$.

Although an M/G/m queuing system has been considered (see, e.g., [8]), the M/M/m queuing model is the only model that accommodates an analytical and closed form expression of the probability density function of the waiting time of a newly arrived service request.

Let be the average service rate, i.e., the average number of service requests that can be finished by a server of S in one unit of time is

$$1/\bar{x} = s/\bar{r}$$

The server utilization is

$$\bar{x}/m \qquad /m.r/s,$$

Which is the average percentage of time that a server of S is busy.

Let p_k denote the probability that there are k service requests (waiting or being processed) in the M/M/m queuing system for S.

$$p_k = \begin{cases} p_0 \dfrac{(m\rho)^k}{k!}, & k \le m; \\ p_0 \dfrac{m^m \rho^k}{m!}, & k \ge m, \end{cases}$$

Where

$$p_0 = \left(\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \cdot \frac{1}{1-\rho} \right)^{-1}.$$

The probability of queuing (i.e., the probability that a newly submitted service request must wait because all servers are busy) is

$$P_q = \sum_{k=m}^{\infty} p_k = \frac{p_m}{1-\rho} = p_0 \frac{(m\rho)^m}{m!} \cdot \frac{1}{1-\rho}.$$

The average number of service requests (in waiting or in execution) in S is

$$\overline{N} = \sum_{k=0}^{\infty} k p_k = m\rho + \frac{\rho}{1-\rho} P_q.$$

Applying Little's result, we get the average task response time as

$$\overline{T} = \frac{\overline{N}}{\lambda} = \bar{x}\left(1 + \frac{P_q}{m(1-\rho)}\right) = \bar{x}\left(1 + \frac{p_m}{m(1-\rho)^2}\right).$$

The average waiting time of a service request is

$$\overline{W} = \overline{T} - \bar{x} = \frac{p_m}{m(1-\rho)^2}\bar{x}.$$

The waiting time is the source of customer dissatisfaction. A service provider should keep the waiting time to a low level

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014

by providing enough servers and/or increasing server speed, and be willing to pay back to a customer in case the waiting time exceeds certain limit.

IV. PROBLEM FORMULATION AND SOLUTION

A networked cloud request is modeled as a weighted undirected graph denoted by $GV(NV, EV)$ where N^V represents the set of virtual nodes and EV the set of virtual links. Similarly, the substrate network is modeled as a weighted undirected graph $GS = (NS, ES)$.

A. Networked Cloud Mapping:

The allocation of physical resources (substrate nodes, links, and paths) to the networked cloud is determined by Resource mapping based on request. Request mapping is comprised of node assignment and link assignment. Specifically, node assignment is denoted as:

$$M^N : N^V \rightarrow N^S$$
$$where\ M^N(n^V) \in V_a^S, n^V \in V_a^V \subseteq N^V.$$

Example:

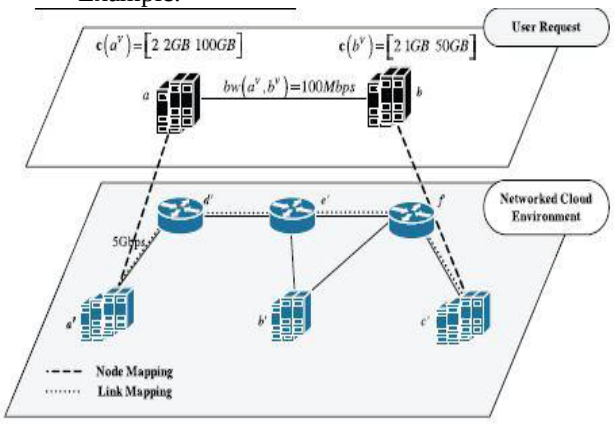

Figure. 1. Networked cloud environment and request mapping

B. Mixed Integer Programming Formulation:

The node and link mapping phase are dependent of each other. To the node and link mapping phase, the methodology proposed in [1] is adopted, without however posing any location constraints on virtual nodes. Specifically the substrate network graph is augmented with the virtual nodes of the request. Every newly added (virtual) node in the augmented substrate graph is connected to every substrate node with infinite bandwidth. Hence, the augmented undirected substrate graph is denoted as $G^{S'} = (N^{S'}, E^{S'})$ where

$$_N S' = N^S \quad N^V \text{ and}$$

$$_E S' = E^S \quad \{(n^V, n^S) |\ n^V \quad N^V,\ n^S \quad N^S$$

Every virtual link (n^V, m^V) E^V with bandwidth requirements $bw(n^V, m^V)$ is considered a commodity in the augmented substrate graph originated at the virtual node $_n V_N S' \quad \setminus N^S$ and ending at the virtual node $_m V_N S' \quad \setminus N^S$

The resource allocation problem in the augmented substrate graph is formulated as a mixed integer programming (MIP) $|E^V|$- commodity flow problem, where the communication demands among theNS0 nodes are specified as a $|N^{S'}| * |N^{S'}|$ demand matrix. For the sake of simplicity superscripts V ,S will be omitted in the following.

C. Variables:

x_{uv}^{nm} :a binary variable set to 1 if there is traffic flow of the virtual link (n,m) E^V routed via the augmented substrate link (u,v) $E^{S'}$

f_{uv}^{nm} :the amount of traffic for the virtual link (n,m) E^V routed over the link (u,v) $E^{S'}$ from u to v.

D. Objective:

$$\min \sum_{uv \in E^S} \sum_{nm \in E^V} C_{uv} f_{uv}^{nm} + \sum_{a \in A} \sum_{nm \in E^V} \sum_{w \in V_a^S \subseteq N^S} \sum_{p \in V_a^V \subseteq N^{S'} \setminus N^S} D_w x_{pw}^{nm} \sum_{i \in I} c_i(p) + \sum_{uv \in E^S} \sum_{nm \in E^V} C_{uv} x_{uv}^{nm}.$$

The aims of the minimization problem is:

a. To minimize the cost of mapping the request into the substrate, as provided by the summation terms (total amount of bandwidth allocated on substrate links that are parts of the substrate paths mapped to the requested virtual links and total amount of computational resources that are allocated to the physical servers mapped to requested virtual nodes). The cost of embedding a networked cloud request corresponds to the sum of substrate resources allocated to that request. Each of these terms multiplied by the corresponding monetary factor can provide the cost of embedding the particular request to the cloud provider resources.

Weights C_{uv} and D_w can be adjusted to balance the load on the substrate links and nodes, respectively. As an example, in [7] the weights C_{uv} and D_w have been set equal to the inverse values of the available bandwidth of the link and the specific node-type available capacity, respectively.

b. To minimize the overall number of hops for a virtual link mapped on a substrate path, according to the summation term (an appropriately defined weight). In the particular case, it has been set equal to C_{uv} to associate the length of the substrate path mapped to the virtual link (m,n) and available capacity of links included in the path, since both are implicitly related to latency.

E. Solution:

In previous section, we have seen the NCM problem which aims to minimize the mapping cost and the overall number of hops. To address this we formulated MIP Problem. The main two problem types that MIP addresses in this field are: 1) network synthesis and 2) resource assignment problems [4].

As a solution, the following methodology is applied. The request is mapped to the networked cloud in two phases: 1) solving the flow allocation problem as was described in the

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014

previous section that results in substrate node mapping; and 2) allocating virtual links to the substrate.

F. Node Mapping Phase:

Due to the nature of the MIP problem presented[14], the optimal fractional solution is computed for the problem's linear programming relaxation of the integer variable x_{uv}^{nm} ,

Which can provide a solution at least as good as the integer one. The relaxed problem can be solved by any suitable linear programming method, in polynomial time (e.g., CPLEX dual simplex routine). A rounding technique is applied to obtain the integer solution of the aforementioned relaxed MIP problem. Randomized rounding for LP relaxations was introduced by Raghavan and Thompson for multicommodity routing problems, where the fractional values contained in the optimal LP solution were treated as probabilities.

The randomized rounding technique proposed in [1] is adopted, where the correlation between the linear variable f_{uv}^{nm} and the binary variable x_{uv}^{nm} , during the LP relaxation process is maintained. Specifically, the substrate node that maximizes the x_{uv}^{nm} , f_{uv}^{nm} product per virtual link is selected.

G. Link Mapping Phase:

Once the aforementioned node mapping procedure has been successfully completed, link mapping is achieved by solving the multi commodity flow allocation problem allowing traffic bifurcation [4]. Alternatively, a shortest path algorithm can be applied to restrict each flow to a single path.

V. PROFIT MAXIMIZATION

To formulate and solve our optimization problems analytically, we need a closed-form expression of C. To this end, let us use the following closed-form approximation,

$$\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} \approx e^{m\rho},$$

Which is very accurate when m is not too small and _ is not too large . We also need Stirling's approximation of m!

i.e., $$m! \approx \sqrt{2\pi m}\left(\frac{m}{e}\right)^m.$$

Therefore, we get the following closed-form approximation of P_m :

$$p_m \approx \frac{1-\rho}{\sqrt{2\pi m}(1-\rho)(e^\rho/e\rho)^m + 1},$$

And the following closed-form approximation of P_q

$$P_q \approx \frac{1}{\sqrt{2\pi m}(1-\rho)(e^\rho/e\rho)^m + 1}.$$

By using the above-closed-form expression of P_q , we get a closed-form approximation of the expected service charge to a service request as

$$C \approx a\bar{r}\left(1 - \frac{1}{(\sqrt{2\pi m}(1-\rho)(e^\rho/e\rho)^m + 1)} \times \frac{1}{((ms-\lambda\bar{r})(c/s_0 - 1/s) + 1)} \times \frac{1}{((ms-\lambda\bar{r})(a/d + c/s_0 - 1/s) + 1)}\right).$$

For convenience, we rewrite C as

$$C = a\bar{r}\left(1 - \frac{1}{D_1 D_2 D_3}\right),$$

where

$$D_1 = \sqrt{2\pi m}(1-\rho)(e^\rho/e\rho)^m + 1,$$
$$D_2 = (ms - \lambda\bar{r})(c/s_0 - 1/s) + 1,$$
$$D_3 = (ms - \lambda\bar{r})(a/d + c/s_0 - 1/s) + 1.$$

Our discussion in this section is based on the above-closed form expression of C[13].

VI. CONCLUSION

Here, we have seen the virtual resource allocation problem for networked cloud environments, incorporating heterogeneous substrate resources, and provide an appropriate approximation approach to address the problem. For the node mapping phase- MIP problem formulation. For the link mapping phase- multicommodity flow problem.

By using an M/M/m queuing model, we formulated and solved the problem of optimal multiserver configuration for profit maximization in a cloud computing environment. Our discussion can be easily extended to other service charge functions and to other pricing models. The main focus is placed on wired and fixed networks which can be extended for dynamic heterogeneous environments (eg: wireless).

VII. REFERENCES

[1] M. Chowdhury, M.R. Rahman, and R. Boutaba, "ViNEYard: Virtual Network Embedding Algorithms With Coordinated Node and Link Mapping," IEEE/ACM Trans. Networking, vol. 20, no. 1, pp. 206-219, Feb. 2012, doi: 10.1109/TNET .2011.2159308.

[2] W. Szeto, Y. Iraqi, and R. Boutaba, "A Multi-Commodity Flow Based Approach to Virtual Network Resource Allocation," Proc. IEEE GLOBECOM '03, vol. 6, pp. 3004-3008, Dec. 2003, doi:10.1109/GLOCOM.2003.1258787.

[3] J. Fan and M.H. Ammar, "Dynamic Topology Configuration in Service Overlay Networks: A Study of Reconfiguration Policies," Proc. IEEE INFOCOM '06, pp. 1-12, Apr. 2006, doi:0.1109/ INFOCOM. 2006.139.

[4] M.G.C. Resende and P. Pardalos, Handbook of Optimization in Telecommunication. Springer, 2006.

[5] M. Armbrust et al., "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report No. UCB/EECS-2009-28, Feb. 2009.

[6] 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 Fifth Utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, 2009.

[7] K. Hwang, G.C. Fox, and J.J. Dongarra, Distributed and Cloud Computing. Morgan Kaufmann, 2012.

[8] H. Khazaei, J. Misic, and V.B. Misic, "Performance

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014

Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 5, pp. 936-943, May 2012.

[9] Y.C. Lee, C. Wang, A.Y. Zomaya, and B.B. Zhou, "Profit-Driven Service Request Scheduling in Clouds," Proc. 10th IEEE/ACM Int'l Conf. Cluster, Cloud and Grid Computing, pp. 15-24, 2010.

[10] K. Li, "Optimal Load Distribution for Multiple Heterogeneous Blade Servers in a Cloud Computing Environment," Proc. 25th IEEE Int'l Parallel and Distributed Processing Symp. Workshops, pp. 943-952, May 2011.

[11] C.S. Yeo and R. Buyya, "A Taxonomy of Market-Based Resource Management Systems for Utility-Driven Cluster Computing," Software - Practice and Experience, vol. 36, pp. 1381-1419, 2006.

[12] J. Lu and J. Turner, "Efficient Mapping of Virtual Networks onto a Shared Substrate," Technical Report WUCSE-2006-35, Washington Univ. St. Louis, 2006.

[13] Junwei Cao, Kai Hwang, Keqin Li, and Zomaya A.Y, "Optimal Multiserver Configuration for Profit Maximization in Cloud Computing," IEEE Trans. Parallel and Distributed Systems, vol. 24, no. 6, pp. 1087-1096, Jun. 2013, doi: 10.1109/TPDS.2012.203.

[14] C. Papagianni, A. Leivadeas, S. Papavassiliou, V. Maglaris, C. Cervello-Pastor, A. Monje, " On the Optimal Allocation of Virtual Resources in Cloud Computing Networks," IEEE Trans. Computers, vol. 62, no. 6, pp. 1060-1071, Jun. 2013, doi: 10.1109/TC.2013.31

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014