International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 10, Issue 3 (March 2014), PP.40-47

Dynamic Consolidation of Virtual Machines In Cloud Data Centers For Managing Overloaded Hosts Under Quality of Service Constraints

P.D.Saronrex[1] R.K.Maheswari[2]
[1]*Trichy Engineering College, Assistant Professor, Department of IT, Somu nagar, Konalai, Trichy, Tamilnadu, India.*
[2] *PG Scholar Trichy Engineering College, Department of CSE, Somu nagar, Trichy, Tamilnadu, India.*

Abstract:- Cloud computing has emerged as the default paradigm for a variety of fields especially considering the resources and infrastructure consumption in case of distributed access. Overloaded host is an aspect of dynamic VM consolidation that directly influences the resource utilization and Quality of Service (QoS) delivered by the system. The server overloads cause resource shortages and performance degradation of applications. Current solutions to the problem of host overload detection are generally heuristic-based, or rely on statistical analysis of historical data. The limitations of these approaches are that they lead to sub-optimal results and do not allow explicit specification of a QoS goal proposing a novel approach that for any known stationary workload and a given state configuration optimally solves the problem of host overload detection by maximizing the mean inter-migration time under the specified QoS goal based on a Markov chain model and proposed a control algorithm for the problem of host overload detection as a part of dynamic VM consolidation. The model allows a system administrator to explicitly set a QoS goal in terms of the OTF parameter, which is a workload independent QoS metric. For a known stationary workload and a given state configuration, the control policy obtained from the Markov model optimally solves the host overload detection problem in the online setting by maximizing the mean inter-migration time, while meeting the QoS goal. Using the Multisize Sliding Window workload estimation approach the model to handle unknown non stationary workloads and propose an optimal offline algorithm for the problem of host overload detection to evaluate the efficiency of the MHOD algorithm.

Keywords:- *Distributed systems, Cloud computing, virtualization, dynamic consolidation, energy efficiency, host overload detection*

I. INTRODUCTION

Cloud computing [1] is a promising computing paradigm which recently has drawn extensive attention from both academia and industry. By combining a set of existing and new techniques from research areas such as Service-Oriented Architectures (SOA) and virtualization, cloud computing is regarded as such a computing paradigm in which resources in the computing infrastructure are provided as services over the Internet. Along with this new paradigm, various business models are developed, which can be described by terminology of "X as a service (XaaS)" where X could be software, hardware, data storage, and etc. Successful examples are Amazon's EC2 and S3 [1], Google App Engine [3], and Microsoft Azure [4] which provide users with scalable resources in the pay-as-you use fashion at relatively low prices. For example, Amazon's S3data storage service just charges $0.12 to $0.15 per gigabyte month. As compared to building their own infrastructures, users are able to save their investments significantly by migrating businesses into the cloud. With the increasing development of cloud computing technologies, it is not hard to imagine that in the near future more and more businesses will be moved into the cloud. As promising as it is, cloud computing is also facing many challenges that, if not well resolved, may impede its fast growth. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Private Cloud uses groups of public or private server pools from an internal corporate data center. It has to be managed by Enterprise, and allows fine grain access to resources .Private Clouds are generally the solution considered by enterprises that do not want to outsource any part of IT infrastructure and services for security concerns. Amplify our datacenter's efficiency and agility while enhancing security and control with a private cloud from VMware Consolidate datacenters and deploy workloads on infrastructure with built-in security and role-based access control. Migrate workloads between pools of infrastructure and integrate existing management systems using customer extensions, APIs, and open cross-cloud standards. Deliver cloud infrastructure on-demand so end users can consume virtual resources with maximum agility.

Cloud Computing [1] have grown in popularity over the past couple of years with this cloud computing services, the shared infrastructure means it works like a utility: we only pay for what we need, upgrades are automatic, and scaling up or down is easy. In order to implement particular resource management techniques such as VM multiplexing or VM live migration that, even if transparent to final users, has to be considered in the design of performance models in order to accurately understand the system behavior. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models. On-demand self-service, broad network access, Resource pooling, Rapid elasticity & Measured Services.

II. RELATED WORK

Prefetching technology is a widely used computer system technology, which hides system expends at every lay of technology, transmission expends. User model [1] is mainly used to collect user information, user status information and user behavioural habit information, then organize and manage this information. Cloud services model [2] can organize and manage service registration information, service current status information and services relevant information. Cloud status model is focus on describing and maintaining current resource status information, periodical performance regulations information and current task executing information. The data obtain mechanism prefetches and analyzes user data, given factors affecting data prefetching hit ratio from the global perspective, it carries out cooperative prefetching analysis and enhances services quality.

Cloud computing is increasingly gaining inroads among a variety of organizational users. As clouds are introduced for use by enterprises, service providers, and governmental and educational entities, new challenges related to the interconnection between such clouds emerge. Cloud administrators seek to maintain acceptable levels of autonomy and control over their cloud infrastructure, while ensuring the integrity of the cloud services. At the same time; they are expected to enable cross-cloud services, including mobility of workloads between clouds.

We present the design and implementation of a technology that enables live mobility of virtual machines between clouds, while enforcing the cloud insularity requirements of autonomy, privacy, and security. We also provide an empirical evaluation of our solution, demonstrating its viability and compliance with requirements. Dynamic resource management [3, 10] it is an active area of research. The authors of employ prediction techniques and queuing theory results to allocate resources efficiently within a single server serving a web workload. Static allocation approach is used in where authors propose a simple heuristic for vector bin-packing problem and apply it to minimize the number of servers required to host a given web traffic. In control theory is applied to design a system for performance control of web server. The arrival rate of requests to the server is throttled based on the feedback system. The authors propose an optimization algorithm that allocates resources (i.e., web servers) depending on the expected financial gain for the hosting center.

III. PROPOSED MODEL

VMware [13] is the global leader in virtualization and cloud infrastructure. VMware offers a unique, evolutionary path to cloud computing that reduces IT complexity, significantly lowers costs and enables more flexible, agile service delivery. VMware vSphere leverages the power of virtualization to transform datacenters into simplified cloud computing infrastructures and enables educational institutions to deliver flexible and reliable IT services. VMware vSphere virtualizes and aggregates the underlying physical hardware resources across multiple systems and provides pools of virtual resources to the datacenter. As a cloud operating system, VMware vSphere manages large collections of infrastructure (such as CPUs, storage, and networking) as a seamless and dynamic operating environment, and also manages the complexity of a datacenter. The following component layers make up VMware vSphere.

Cloud based systems are inherently large scale, distributed, almost always virtualized, and operate in automated shared environments. Performance and availability of such systems are affected by a large number of parameters including characteristics of the physical infrastructure (e.g., number of servers, number of cores per server, amount of RAM and local storage per server, configuration of physical servers, network configuration, persistent storage configuration), characteristics of the virtualization infrastructure (e.g., VM placement and VM resource allocation, deployment and runtime overheads), failure characteristics (e.g., failure rates, repair rates, modes of recovery), characteristics of automation tools used to manage the cloud system, and so on proposing a novel approach that for any known stationary workload and a given state configuration optimally solves the problem of host overload detection by maximizing the mean inter-migration time based on a Markov chain model and proposed a control algorithm for the problem of host overload detection as a part of dynamic VM consolidation, optimally solves the host overload detection problem in the online setting by maximizing the mean inter-migration time.

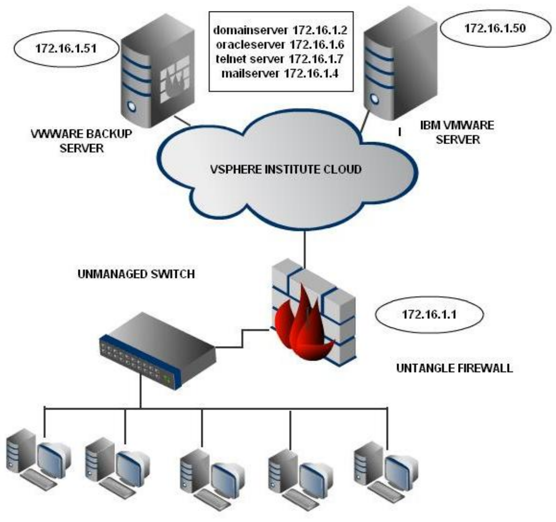

Fig 3.0: End User Cloud

Infrastructure Services: Infrastructure Services are the set of services provided to abstract, aggregate, and allocate hardware or infrastructure resources. Infrastructure Services are categorized into several types.

- VMware vCompute, which includes the VMware capabilities that abstract away from underlying disparate server resources. vCompute services aggregate these resources across many discrete servers and assign them to applications.
- VMware vStorage, which is the set of technologies that enables the most efficient use and management of storage in virtual environments.
- VMware vNetwork, which is the set of technologies that simplify and enhance networking in virtual environments.

Application Services: Application Services are the set of services provided to ensure availability, security, and scalability for applications. Examples include High Availability and Fault Tolerance.

VMware vCenter Server: Provides a single point of control of the datacenter. It provides essential datacenter services such as access control, performance monitoring, and configuration.

Clients: Users can access the VMware vSphere datacenter through clients such as the vSphere Client or Web Access through a Web browser.

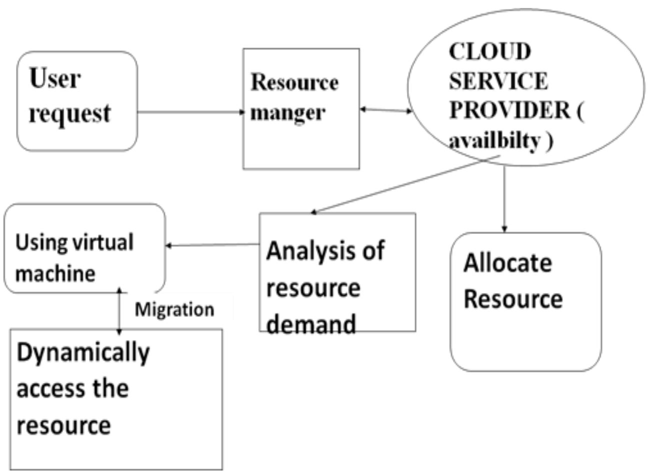

Fig 3.0: Overall architecture

3.1 Dynamic Resource Allocation

Fig 3.1: Resource allocations

Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed accurately. Resource provisioning solves that problem by allowing the service providers to manage the resources for each individual module.

3.2 Cloud Service Provider

Fig: 3.2 service provider

The cloud service provider is responsible for maintaining an agreed-on level of service and provisions resources accordingly. A CSP, who has significant resources and expertise in building and managing distributed cloud storage servers, owns and operates live Cloud Computing systems, it is the central entity of cloud. Cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job. The order and time of allocation of resources are also an input for an optimal resource allocation.

3.3 Cloud Consumer

Cloud consumer represents a person or organization that maintains a business relationship with, and uses the service from, a cloud provider. Users, who stores data in the cloud and rely on the cloud for data computation, Cloud consists of both individual consumers and organizations. Cloud consumers use Service-Level Agreements (SLAs) for specifying the technical performance requirements to be fulfilled by a cloud provider. A cloud provider may also list in the SLAs a set of restrictions or limitations, and obligations that cloud consumers must accept.

3.4 Performance Evaluation

In cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. Some of the strategies discussed above mainly focus on CPU, memory resources .secured optimal resource allocation algorithms and framework to strengthen the cloud computing paradigm.

IV. METHODOLOGY

VMware [13] has successfully implemented dozens of private cloud infrastructures. To help we leverage the experience and best practices we have accumulated from these deployments, we have developed the Vcloud and Vsphere of our institute architecture-a set of documents we can use to better understand both the principles upon which VMware's cloud strategy is executed, and the mechanics for us to implement our own cloud infrastructure.

4.1 Algorithm

4.1.1 An Optimal Offline Algorithm

Input: A system state history
Input: M, the maximum allowed OTF
Output: A VM migration time

1: **while** history is not empty **do**
2: **if** OTF of history _ M **then**
3: **return** the time of the last history state
4: **else**
5: drop the last state from history
6: **end if**
7: **end while**

4.1.2 MHOD Algorithm

Input: A CPU utilization history
Output: A decision on whether to migrate a VM
1: **if** the CPU utilization history size > Tl **then**
2: Convert the last CPU utilization value to a state
3: Invoke the Multisize Sliding Window estimation to obtain the estimates of transition probabilities
4: Invoke the MHOD-OPT algorithm
5: **return** the decision returned by MHOD-OPT
6: **end if**
7: **return false**

4.1.3 Window Size Selection Algorithm

Input: J, D, NJ , t, i, j
Output: The selected window size
1: lw   J
2: **for** k = 0 to NJ □ 1 **do**
3: **if** S(i; j; t; k) _ Vac(bpij(t; k); k) **then**
4: lw   J + kD
5: **else**
6: **break loop**
7: **end if**
8: **end for**
9: **return** lw

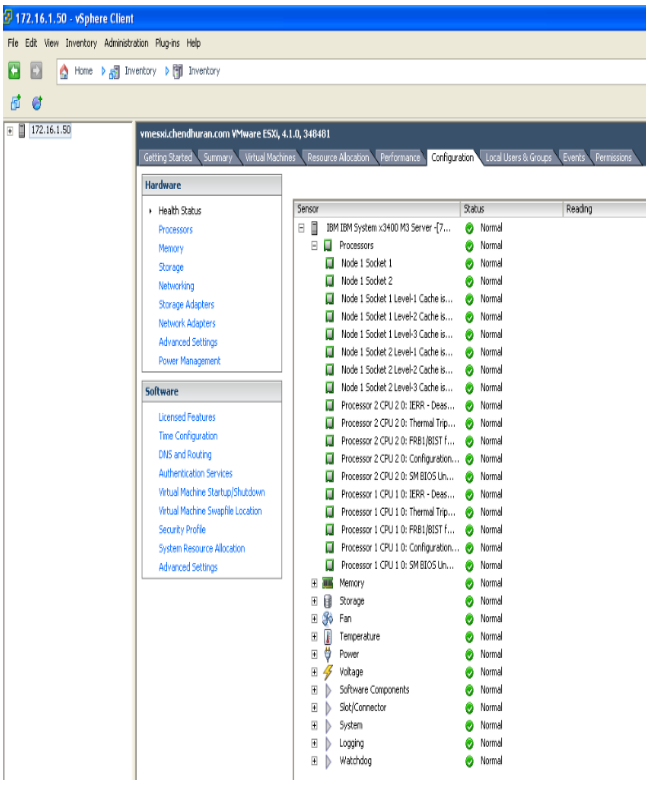

Fig.4.1: Institute vsphere config

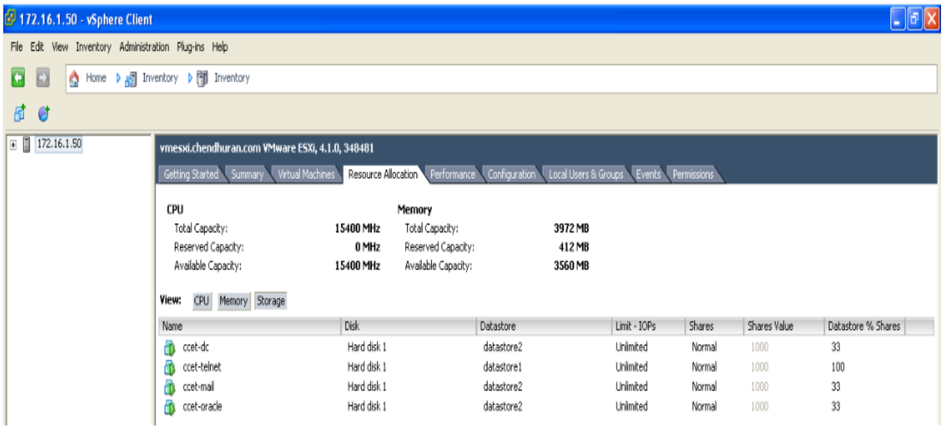

Fig.4.2: Institute data storage

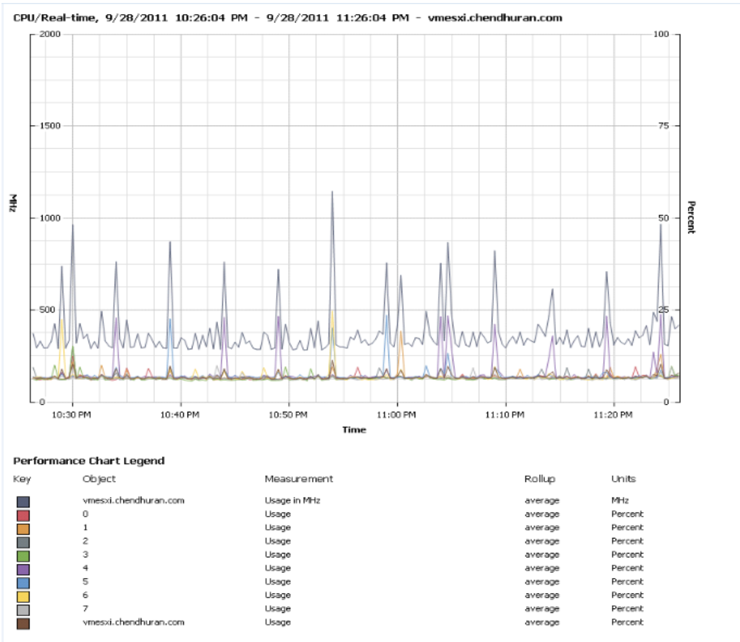

Fig 4.3: Institute hype

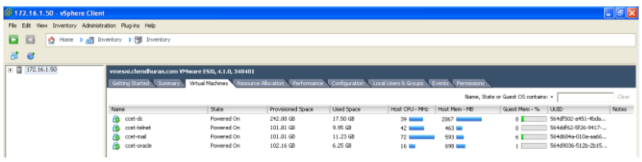

Fig 4.4: Institute virtual servers

Comprehensive Solutions:

- Enable the management of all project and institutional documents.
- Provides comprehensive, secured storage and backup facility of personal and work data.
- Facilitates the exchange of information within across educational institutions.

Flexibility adapting to the way we work:

- Offers the flexibility o f uploading unrestricted format types securely without much hassle.
- Multi-level access control – adapt to an educational institution's hierarchy.
- Allows institution entities to host the entire application on our own servers.

Sharing information:

- Allows sharing of files internally.
- Able to control the permission to allow third party to view and download.

Manage files online:

- Able to organize files in different folders as desired and being able to do it online.
- Single storage and backup repository of all important files.
- Ease of use and user friendly interface.

Security and Accessibility:

- Able to access the file anywhere, anytime within the institution.
- It will be more secure data transmission while number of users accessing the data is scalable.

V. CONCLUSION

Virtualization in computing is the creation of a virtual, rather than actual version of a storage device or network resources. Using some interfaces we can access the data in cloud. This paper gives about the cloud data management interface by using storage virtualization mechanism. The open cloud computing interface is an emerging standard for interoperable interface management in the cloud. Cloud computing can solve complex set of tasks in shorter time by proper resource utilization. To make the cloud to work efficiently, best resource allocation strategies have to be employed. Utilization of resources is one of the most important tasks in cloud computing environment the various strategies have been studied and classified. The different features of the algorithms have been studied. .This can be extended to models which represent PaaS and SaaS Cloud systems and to integrate the mechanisms needed to capture VM migration and data center consolidation aspects that cover a crucial role in energy saving policies.

As a part of future work, a plan to implement the MHOD algorithm as an extension of the VM manager within the OpenStack Cloud platform[7] to evaluate the algorithm in a real system as a part of energy-efficient dynamic VM consolidation. The purpose of this paper is to develop a Private Cloud for Educational firm which would like to do their automation without spending lot on Infrastructure. Forming an Enterprise Cloud where the entire request will reach at set of Server (IBM Cloud Servers) for computation & resource access and hence providing service over the Cloud. It means unnecessarily there is no need to have 100 computers for 32 minutes of work, instead have 1 cloud for 300 different tasks. In such case since everything will be maintained by the institution firm security issues will be of less concern since Private Cloud Infrastructure is being formed internally by the firm. Using VMware ESXi Container, we can improve Simple Network Framework to a Private Cloud for different institutes.

REFERENCES

[1] Amazon Web Services (AWS), Online at http://aws. amazon.com.

[2] Microsoft Azure, http://www.microsoft.com/azure/.

[3] Google App Engine, Online at http://code.google.com/appengine/.

[4] Brock, M.; Goscinski, A.; Grids vs. Clouds Future Information Technology (FutureTech), 2010 5th IEEE International Conference2010 pp 1-6.

[5] Daryl C. Plummer, Thomas J. Bittman, Tom Austin, David W. Cearley, David Mitchell Smith "Cloud Computing: Defining and Describing an Emerging Phenomenon".

[6] High-Performance Cloud Computing: A View of ScientificApplications by Christian Vecchiola, Suraj Pandey, and Rajkumar Buyya.

[7] Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities Buyya, R. Chee Shin Yeo Venugopal, S.

[8] Performance Evaluation of Cloud Computing Offerings Vladimir Stantchev, SOA and Public Services Research Group, TU Berlin.

[9] Schaper, J Cloud Services Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference 2010 pp 91-92.

[10] Srinivasa, K.G.; Siddesh, G.M.; Cherian, S.;" Fault-Tolerant Middleware for Grid Computing "Srinivasa, K.G. Siddesh, G.M Cherian, S.High Performance Computing and Communications (HPCC), 2010. 12th IEEE International Conference pp 635 – 640.

[11] Xingchen Chu Nadiminti, K. Chao Jin Venugopal, S.Buyya, R, Univ. of Melbourne, Melbourne "Aneka: Next- Generation Enterprise Grid Platform for e-Science and e- Business Applications".

[12] Zhao, Peng; Huang, Ting-lei; Liu, Cai-xia; Wang, Xin; Research of P2P architecture based on cloud computing IEEE International Conference.

[13] http://www.vmware.com/files/pdf/techpaper/VMW-TWP-vSPHR-SECRTY-HRDNG-USLET-101-WEB-1.pdf.

[14] Http://www.vmware.com/files/pdf/vc_dbviews_40.pdf.

[15] R. Buyya et al., "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Gener. Comput. Syst., vol. 25, pp. 599–616, June 2009.

[16] "A Iosup-2010" on the performance variability of production cloud services.
[17] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing SLA violations," in Proc. of the 10th IFIP/IEEE Intl. Symp. on Integrated Network Management (IM), 2007, pp. 119–128.
[18] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," Concurrency and Computation: Practice and Experience (CCPE), 2012, DOI: 10.1002/cpe.1867, (in press).
[19] C Strack - 2012 "Performance and Power Management for Cloud Infrastructures".
[20] C Tang-2007" A Scalable Application Placement Controller for Enterprise Data Centers.
[21] D. Gmach, J. Rolia, L. Cherkasova, G. Belrose, T. Turicchi, and A. Kemper, "An integrated approach to resource pool management: Policies, efficiency and quality metrics," in Proc. of the 38th IEEE Intl. Conf. on Dependable Systems and Networks (DSN), 2008, pp. 326–335.
[22] D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper, "Resource pool management: Reactive versus proactive or lets be friends," Computer Networks, vol. 53, no. 17, pp. 2905–2922, 2009.