SURVEY OF VARIOUS SERVER CONSOLIDATION SCHEMES IN CLOUD COMPUTING ENVIRNMENT

[1]Parul H Vaghamshi, [2]Dr. Dhaval A. Parikh
[1]ME student, [2]Head of Department
[1]Department of Computer Engineering
[1]L. D. College of Engineering, Gujarat Technological University, Ahmedabad, India

Abstract : Due to over reliance of today's cloud computing and high performance computing, the size and amount of the data center develop rapidly , which has led to an essential increase in data production, network traffic and energy consumption. As a result energy and operation cost of data centers has brought remarkable challenges in field of cloud computing. When a user application is deployed in the cloud, depending on the SLA, cloud service provider deploys servers for smooth running of the user application. The total number of servers deployed for a user application must be optimal, because underutilized servers are not economical for both cloud service provider and cloud user. Underutilized servers consume power when they are idle; hence deploying optimal number of servers is critical in the operation of the cloud. Server consolidation is a popular approach to reduce cloud data center's energy consumption by minimizing the number of active physical machines. Main focus of research will be on optimizing resource utilization by server consolidation considering parameter like performance and power consumption.

***Index Terms* – Cloud Computing ,Server Consolidation ,VM Consolidation ,energy efficiency.**

I. INTRODUCTION

Cloud computing is shared pools of configurable computer system resources and higher-level services that can be rapidly provisioned with minimal management effort, often over the Internet. Cloud computing relies on sharing of resources to achieve coherence and economies of scale, similar to a public utility. Cloud computing is a new paradigm which deliver computing resources as utility. Datacenters as cloud infrastructure encounter with several issues such as power management for the sake of economic viewpoint. Cloud providers try to find a way to reduce their overall costs. The total cost of ownership (TCO) includes fixed costs, capital expenditure, and variable costs, operational expenditures. Lager amount of operational expenditure is related to server sprawl. Server sprawl is a phenomenon which resources are dispersed through systems with low utilization making high rate of power consumption.

II. CLOUD COMPUTING AND ENERGY USE

Cloud computing services are powered by large datacenters comprising numerous virtualized server instances, high-bandwidth networks, and supporting systems such as cooling and power supply. The equipment can be classified into hardware and software, which are accessed by remote users (see Figure 1)[1]. In terms of hardware, users access cloud services through network equipment that connects servers to the Internet. User software, or appliances, run on top of servers and are managed by a cloud management system (CMS). Other supporting equipment is beyond this article's scope, but typically includes power supply, cooling, and the datacenter building itself.

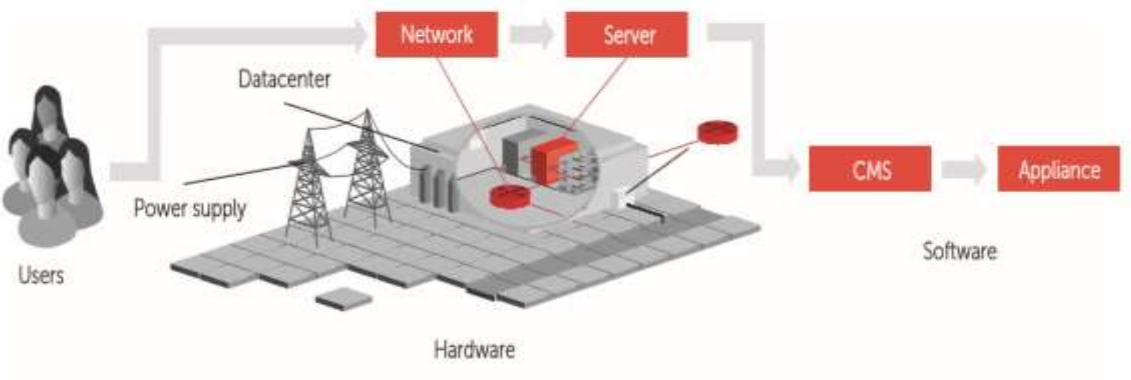

Figure 1 Cloud computing datacenter domains. Equipment includes hardware—such as network equipment and servers—and software such as a cloud management system (CMS) and appliances (that is, user software).

Energy efficiency can be defined as a reduction of energy used for a given service or level of activity.4 However, because of a datacenter equipment's scale and complexity, it's extremely difficult to define a unique service or activity that could be examined for its energy efficiency. Therefore, as Figure 2[1] shows, we identify four scenarios within a system in which energy is used inefficiently—that is, it's lost or wasted.

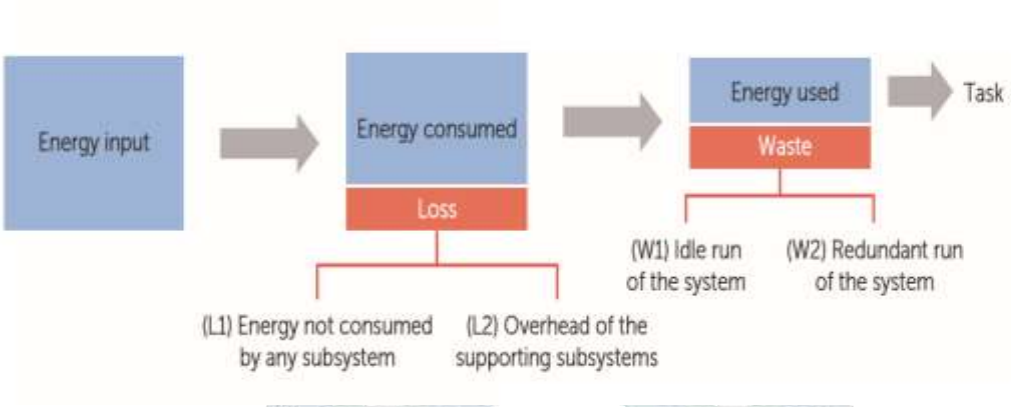

Figure 2 Energy inefficiency. Four scenarios in which a system might lose or waste energy.

Both "loss" and "waste" define inefficient energy usage from an agnostic point of view. Energy loss refers to an energy brought to the system but not consumed by any of its subsystems (L1), such as when energy is lost due to transport or conversion. Loss also includes energy overhead of the supporting subsystems (L2), such as cooling or lighting within a datacenter that mainly provisions cloud services. Energy waste refers to energy used for its primary purpose but wasted due to an idle run of the system (W1), such as when the processor is turned on but running idle. Another source of energy waste is a redundant run of the system (W2), such as keeping a cooling system at maximum during the night, when temperatures are lower.

III. SERVER CONSOLIDATION

Consolidation is an approach to the efficient usage of (physical) servers in order to reduce the total number of servers that an organization requires. The practice developed in response to the above-described server sprawl, a situation in which multiple underutilized servers take up more space and consume more energy than can be justified by their workload [2]. Server virtualization provides technical means to consolidate multiple servers leading to increased utilization of physical servers [3]. A significant chance for power optimization will be presented by consolidation of applications in cloud computing environment. There are significant inter-relationships between resource utilization, performance of consolidated workloads and power consumption. Clusters in datacenters in idle or low utilization status consume large amount of electricity as energy. For example, the energy consumption of non-operative, but turned on, accounts for approximately 70% of the full loaded server energy consumption [5].

Virtualization technique is widely used in cloud datacenter allowing different virtual machines co-host on same physical machine; it also applies consolidation approach to pack virtual machines over minimum number of physical machines [4]. QoS-aware schemes is very important in this ambit because awkward consolidation algorithms neglecting migration cost and applications affinity on special resources may nullify the benefit of consolidation yielding high SLA violation rate. To deploy the requested jobs in cloud environment, the user makes a request to a resource broker, specifying the number of processing units required and the associated memory requirements. If the requested CPU and memory resources are available, the job is accepted. This static strategy ensures that all jobs accepted into the cluster will have sufficient processing units and memory to complete their work. Nevertheless, it can lead to a waste of resources, as many workloads proceed in phases, not all of which use all of the allocated processing units at all times. Consolidation technique dynamically declines number of active servers by releasing unnecessary machines in the current computing phase [6]. Fig. 2. Illustrates consolidation scheme.

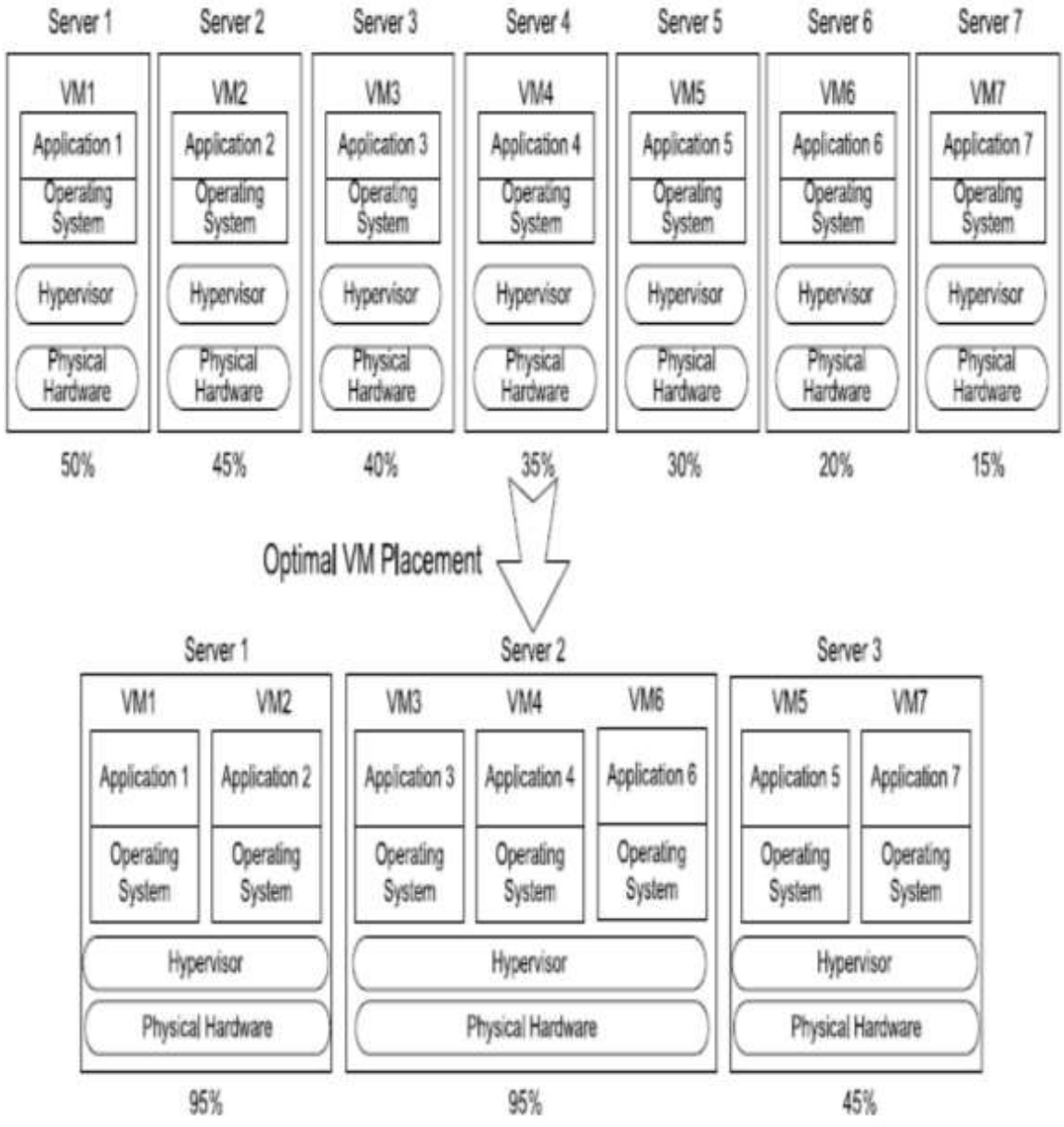

Figure 3 VMs Consolidation Scheme [7].

After server consolidation by VMs migration, servers numbered 4 through 7 will be set into lower power consumption state or hibernate mode to save energy [7,2]. Server consolidation is tantalizing approach, because effective consolidation is not as trivial as packing the maximum workload in the smaller number of servers with keeping each resource (CPU, Disk, Network, etc.) on every server near 100% utilization. In practice, at least two concerns makes consolidation problem become very hard decision making problem. Firstly, consolidation methods must carefully decide which workload should be combined on the same physical server. In fact, understanding the nature of workload and their affinity to resources are crucial. Secondly, the migration cost in terms of additional resource usage and migration time may lead performance degradation and prolongs workload execution. Therefore, the consolidation problem is more complex than bin packing problem. In literature, some researchers abstracts server consolidation problem as multi-dimensional bin packing problem where servers are bins with each resource (CPU, disk, network, etc.) being one dimension of the bin. The bin size along each dimension is given by the energy optimal utilization level. Each hosted application with known resource utilizations can be treated as an object with given size in each dimension. Minimizing the number of bins should minimize the idle power wastage [8]. However, that is not true in general, according to aforementioned reasons causing the energy aware consolidation problem to differ from traditional vector bin packing.

IV. A STUDY OVER SERVER CONSOLIDATION SCHEMES IN LITERATURE

Benjamin et al. in [3] have studied mathematical programming approaches for server consolidation problems in virtualized large scale data centers. They have shown aforementioned problem is strongly NP-Hard and assumed it is multidimensional bin packing problem [9]. The authors formulated problem as static server allocation problem in large datacenters and presented a heuristic method called SSAPv, static server allocation problem with variable workload. They considered two widespread type of services being used by users in cloud environment (i) W/A/B services, namely, web, application and database services respectively and (ii) ERP services to show the working of their proposed method.

In [4], a VM consolidation approach has been applied on OpenStack, an open-source platform for Cloud computing to show effectiveness of VM consolidation on power consumption reduction. The proposed approach makes decision to map VMs into the minimal number of physical servers to reduce the runtime power consumption. On the other hand, server consolidation technique is severely resource intensive which may cause service degradation. So, proposed algorithm considers CPU, network and other resource features to avoid performance degradation and SLA violation. Its experimental results show the effectiveness [10].

A network-aware VM consolidation algorithm called interand-intra datacenter VM-Placement has been proposed by burak et al. in [11] to aim energy savings in large data centers in which include multiple medium size datacenters geographically distributed and connected via backbone network. They have formulated VMs consolidation by new mixed integer linear programming (MILP) to place VMs to appropriate hosts by considering users requirement expressed in SLA and CPU, memory, network bandwidth, delay and the distance of users to be hosted the selected datacenter. Their results show the improvements in terms of power consumption, resource utilization and fairness for many connected medium size datacenters.

Authors in [12] have propounded server consolidation algorithm in which their objectives are application performance and system utilization. It leverages virtualization to run away from server sprawl in case of increasing server equipment in demanding IT service rising in which server sprawl makes low server utilization and high system management cost. Current paper uses key performance metrics in their monitoring framework that triggers to consolidate VMs into minimum number of physical hosts. Also, it considers migration cost not to have negative affection on user application performance such as SLA violation.

A linear program and heuristic algorithm has been presented in [13] that controls the number of migration and reduces server energy consumption. It is cost-aware and migration controller algorithm because precludes unnecessary migrations due to unpredictable workloads that require VM resizing. The algorithm prioritizes VMs with steady capacity to obstruct residue migrations. The authors have formulated it as linear programming problem. Their heuristics a little improves the famous first-fit decreasing (FFD), best-fit decreasing (BFD), worst-fit decreasing (WFD), and almost worst-fit decreasing (AWFD) algorithms [14].

V. CONCLUSION AND FUTURE WORK

Power usage is the first class concern in power hungry datacenters regard to economic viewpoint. The main reason causes power wastage is related to low resource utilization. So, to deal with the problem several works have been proposed with consolidation approach. After studying over published paper, we have presented our comparison framework. Then we reviewed papers and finally compared based on our subjective comparison framework. Server consolidation problem is abstracted to well-known NP-Hard bin-packing problem which items must be packed into minimum number of bins. Therefore, miscellaneous combinatorial algorithms have been propounded to figure out the aforementioned problem. One of the biggest challenges is related to not pay attention user QoS in this environment because the schemes which executes combinatorial algorithm take long time making high rate of user SLA violation. On the other hand, big portion of research work on CPU utilization as only resources which nullify server consolidation schemes in real conditions especially when there is not any meaningful relation between CPU usage and other resources. Future direction can be toward solving open issues such as considering resource vector utilization instead of considering limited number of resources along with development of new combinatorial algorithm to make fast decision in cloud fragile environment.

REFERENCES

[1] Toni Mastelic and Ivona Brandic**:** Recent Trends in Energy-Effi cient Cloud Computing, Vienna University of Technology

[2] L. Jian-ping, X.Li, C. Min-rong, Hybrid shuffled frog leaping algorithm for energy-efficient dynamic consolidation of virtual machines in cloud data centers, Expert Systems With Applications, in press.

[3] B. Speitkamp, M. Bichler, A mathematical programming approach for server consolidation problems in virtualized data,IEEE Tran. Service Comput(2010)266-278

[4] A. Corradi, M. Fanelli, L. Foschini, VM consolidation: A real case based on OpenStack Cloud, Future Generation Computer Systems**,** 32(2014) 118-127

[5] Kusic, D., Kephart, J. O., Hanson, J. E., Kandasamy, N., & Jiang, G. (2009). Power and performance management of virtualized computing environment via lookahead control **.** Cluster Computing ,12(1), 1-15 ,975

[6] Power Efficient VM Consolidation using Live migration-A step towards Green computing: A white paper
[7] Y. Gao, H. Guan, Z. Qi, Y. Hou, L. Liu, A multi-objective ant colony system algorithm for virtual machine placement in cloud computing , Journal of computer system and science,In press
[8] A. Blaglazov et al., A taxonomy and Survey of Energy-Efficient Data Centers and Cloud computing Systems. White paper
[9] M. Garey and R. Graham, "Resource Constrained Scheduling as Generalized Bin Packing," J. Combinatorial Theory, Series A, vol. 21,pp. 257-298, Nov. 1976.
[10] S. Cash et al., Managed infrastructure with IBM Cloud OpenStack Service, Published in: IBM Journal of Research Development (Volume: 60, Issue: 2-3, March-May 2016)
[11] B. Kartarciet. Al., Inter-and-Intra Data Center VM-Placement for Energy-Efficient Large-Scale Cloud Systems, First International workshop on Management and Security technologies for Cloud Computing 2012.
[12] G.Khanna, K. Beaty, G. Kar, A. Kochut, Application Performance Management in Virtualized Server Environments, Network Operations and Management Symposium , 2006. NOMS (2006).
[13] Y. Gao, H. Guan, Z. Qi, Y. Hou, L. Liu, A multi-objective ant colony system algorithm for virtual machine placement in cloud computing, Journal of Computer and System Sciences, In press.
[14] L.T. Kou, G. Markowsky, Multidimensional bin packing algorithms, IBM Journal of Research and Development 21 (5) 1977.