A Study on Energy Efficient Server Consolidation Heuristics for Virtualized Cloud Environment

Susheel Thakur, Arvind Kalia, Jawahar Thakur

Abstract— **With a growing concern over the energy consumption costs by data centers, research efforts are aiming towards energy efficient data centers to maximize the energy efficiency. Server virtualization is emerging as a vital approach towards the consolidation of applications from multiple applications to one server, with a goal to save energy usage. Savings can be achieved by dynamic consolidation with live migration of VMs depending on the resource utilization, virtual network topologies and thermal state of the computing machines. However, little understanding has been obtained about the potential overhead in the energy consumption for virtualized servers in data centers. This paper presents a study of the research work and innovations done by researchers and academicians based on energy-aware dynamic VM consolidation from one host to another in cloud data centers.**

***Index Terms*— Cloud Computing, Quality of Service (QoS), Bin Packing Problem, Constraint Programming.**

I. INTRODUCTION

In 1969, Leonard Kleinrock [15], one of the chief scientists of the original Advanced Research Projects Agency Network (ARPANET) which seeded the Internet, said: *"As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of computer utilities" which, like present electric and telephone utilities, will service individual homes and offices across the country."* This vision of computing utilities based on a service provisioning model anticipated the massive transformation of the entire computing industry in the 21st century whereby computing services will be readily available on demand, like other utility services available in today's society. Similarly, users need to pay providers only when they access the computing services. In addition, users no longer need to invest heavily or encounter difficulties in building and maintaining complex IT infrastructure.

Manuscript received April, 2014.
***Susheel Thakur**, Department of Computer Science, Himachal Pradesh University, Shimla, India,*
***Arvind Kalia**, Department of Computer Science, Himachal Pradesh University, Shimla, India,*
***Jawahar Thakur**, Department of Computer Science, Himachal Pradesh University, Shimla, India,*

Cloud Computing is defined by NIST [18] *as a model for enabling convenient, on demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.* For simplicity, a cloud is a pool of physical computing resources i.e. a set of hardware, processors, memory, storage, networks, etc. which can be provisioned on demand into services that can grow or shrink in real-time scenario [27]. Cloud computing can be classified as a new paradigm for the dynamic provisioning of computing services supported by state-of-the-art data centers that usually employ virtual machines technologies for consolidation and environment isolation purposes [20]. Modern data centers, operating under the cloud computing model are hosting a variety of applications ranging from those that run for a few seconds (e.g. serving requests of web applications such as e-commerce and social networks portals with transient workloads) to those that run for longer periods of time (e.g. simulations or large data set processing) on shared hardware platforms. The need to manage multiple applications in a data center creates the challenge of on-demand resource provisioning and allocation in response to time-varying workloads. Normally, data center resources are statically allocated to applications, based on peak load characteristics, in order to maintain isolation and provide performance guarantees. Until recently, high performance has been the sole concern in data center deployments and this demand has been fulfilled without paying much attention to energy consumption. The fact that power consumption is set to rise 76% from 2007 to 2030[28] with data centers contributing an important portion of this hike emphasize the importance of reducing energy consumptions in the cloud data centers. According to the Gartner report, the average data centers consumes as much as energy as 25,000 households [8]. As energy costs are increasing while availability dwindles, there is a need to shift focus from optimizing data center resource management for pure performance to optimizing for energy efficiency while maintaining high service level performance.

"The total estimated energy level bill for data centers in 2010 is $11.5 billion and energy costs in a typical data centers double every five years ", according to Mckinsey report" [11].

Minimizing the energy consumption or going for Green computing [17], [24] is a major issue of concern. Energy consumed by servers and in cooling data centers of the cloud system is an expensive affair. The United States Environment Protection Agency (EPA) in its report says that energy consumption of only federated servers and data centers in this nation was 100 billion KWh in 2011 and infrastructure and energy (I&E) cost will be 75 percent of the total operation

cost in 2014[21]. Energy consumption of data centers has risen 56 percent from 2005 to 2010 worldwide, and in 2010 it is accounted to be between 1.1 percent and 1.5 percent of the total electricity used[12]. Thus, minimizing energy consumption is important and designing energy efficient data centers has recently received considerable attention of research community.

There are numerous technologies, services, and infrastructure-level configurations that can be used to make cloud computing energy efficient. Virtualization in cloud computing provides a mechanism the hardware and the system resources from a given operating system. This is typically performed within cloud environment across a large set of servers using hypervisor or virtual machine monitor (VMM) that acts as a bridge between the hardware and the operating system (OS) [20]. The cloud computing middleware is deployed on the top of the virtualization technologies to exploit the capability to its maximum potential while maintaining the Quality of service (QoS) to clients.

The concept of consolidation of virtual machines (VMs) is applied to minimize energy consumption as it significantly reduces the percentage of ideal power in the overall infrastructure. Such a consolidation is done either statically or dynamically at run time. In the static approach, the mapping of the VMs to physical machines cannot be changed at the runtime. While dynamic consolidation of VMs allows the reallocation of physical resources at the runtime, when the load on the virtual machines increases or decreases. When there is a low load on the VMs, less physical infrastructure need to be employed to provide certain performance level. And if the load on virtual machine increases, more physical machines can be allocated. The VMs can be migrated to another physical host if the current host gets overloaded. A dynamic consolidation of virtual machines mandates the cloud service provider to monitor the resource utilization of virtual machines, in order to determine how many physical resources have to be assigned for a particular scenario.

Dynamic VM consolidation consists of two basic processes: Migrations of VMs from underutilized hosts to minimize the number of active hosts, and offloading VMs from the physical hosts when they become overloaded to avoid performance degradation as experienced by the VMs. The idle hosts automatically switch to a low-power mode to eliminate the static power and reduce the overall energy consumption by the system. When required, the physical hosts are reactivated to accommodate new VMs or VMs being migrated.

Another major capability provided by virtualization is live machine migration in which VMs between physical servers can be transferred with a close to zero downtime. Using live machine migration, VMs can be dynamically consolidated to leverage the fluctuations in the workload and keep the number of active physical servers at the minimum at all times [5].

Consolidation will result in reduced power consumption and thus reducing overall operational costs for data center administrators. Live migrations can be used to achieve this. Based on the load conditions, under-utilized machines having resource usage above a certain threshold are identified and migrations are triggered to tightly pack VMs to increase overall resource usage on all PMs and try to minimize the overall energy consumption as possible [1].

II. Energy Efficient Server Consolidation Heuristics

Server consolidation is an approach for the efficient usage of computer server resources in order to reduce the total number of servers or server locations that an organization requires. This approach was developed in response to the problem of "server sprawl". In order to reduce server sprawl in the data centers, server consolidation algorithms are implemented. These algorithms are VM packing heuristics which try to pack as many VMs as possible on the physical machine (PM) so that resource usage is improved and under-utilized machines can be turned off, aiming to minimize the energy consumption and enhancing the throughput in the cloud data centers.

A. Sandpiper: Black box and Gray-box resource management for Virtual Machines

Sandpiper is a system that automates the task of monitoring and detecting hotspots, determining a new mapping of physical resources to virtual resources, by resizing or migrating VM's to eliminate the hotspots. Sandpiper makes use of automated black-box and gray box strategies for virtual machine provisioning in cloud data centers. Specifically the black-box strategy can make decisions by simply observing each virtual machine from the outside and without any knowledge of the application resident within each VM. The authors present a gray-box approach that assumes access to OS-level statistics in addition to external observations to better inform the provisioning algorithm. Sandpiper implements a hotspot detection algorithm that determines when to resize or migrate virtual machines, and a hotspot migration algorithm that determines what and where to migrate and how many resources to allocate. The hotspot detection component employs a monitoring and profiling engine that gathers usage statistics on various virtual and physical servers and constructs profiles of resource usage. These profiles are used in conjunction with prediction techniques to detect hotspots in the system. Upon detection, Sandpiper grants additional resources to overloaded servers if available. If necessary, Sandpiper's migration is invoked for further hotspot mitigation. The migration manager employs provisioning techniques to determine the resource needs of overloaded VMs to underloaded servers.

Sandpiper supports both black-box and gray-box monitoring techniques that are combined with profile generation tools to detect hotspots and predict VM Resource requirements. Hotspots are detected when CPU usage values are violated with respect to the CPU thresholds set. Physical machines (PMs) are classified as underloaded or overloaded. The PMs are sorted based on the descending order of their volume metric, and VMs are sorted based on the descending order of their vsr metric, where volume (vol) and vsr are computed as:

$$\text{vol} = \left(\frac{1}{1-\text{cpu}}\right) * \left(\frac{1}{1-\text{mem}}\right) * \left(\frac{1}{1-\text{net}}\right) \qquad (1)$$

$$\text{vsr} = \frac{\text{vol}}{\text{size}} \qquad (2)$$

where cpu, mem and net refers to cpu, memory and n/w usages of the PMs and VMs respectively and size refers to the memory footprint of the VM. vsr is the volume-size metric.

To mitigate hotspot on an overloaded PM, the highest vsr VM is migrated to a least loaded PM amongst the underloaded ones. If the least loaded PM can't house the PM, next PM in the sorted order is checked. Similarly, if the VM cannot be

housed in any of the underloaded PMs, next VM in the sorted order is checked. This way sandpiper tries to eliminate hotspots by remapping VMs on PMs through migration. The experimental results showed that migration overhead is less than that of swapping overhead; however, swapping increases the chances of mitigating hotspots in cluster with high average utilization [23], [25].

B. EnaCloud: An Energy Saving Application

Bo Li et. al proposed EnaCloud which supports application scheduling and live migration to minimize the number of running physical machines in order to save energy. It also aims to reduce the number of migrations of virtual machines. In EnaCloud, there is a central global controller which runs the Concentration Manager and Job Scheduler. The job scheduler receives the arrival of workload, departure and resizing events and deliver them to the Concentration Manager. Then Concentration Manager then generates a series of insertion and migration operations for application placement scheme which are passed to the Job Scheduler. The Job Scheduler then dispatches them to the virtual machine controller by decomposing the schemes.
Each resource node consists of Virtual Machine Controller, Resource Provision Manager and Performance Monitor. The Virtual Machine Controller invokes the hypervisor to execute the commands such as VM start, stop or migrate. The Resource Provision Manager does the resizing of VMs based on the performance statistics collected by the Performance Monitor.

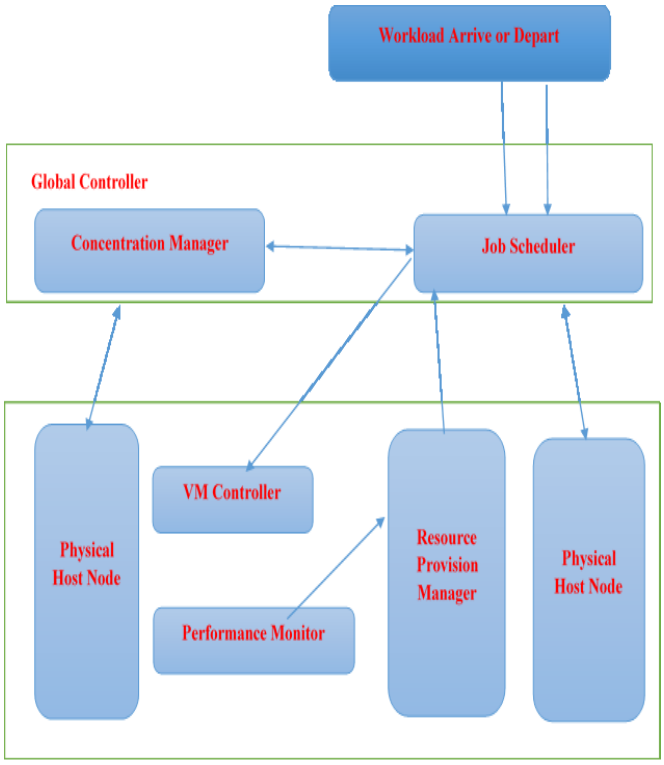


Figure 2.1 EnaCloud Architecture

Here, two types of nodes are considered: computing nodes and storage nodes. Storage nodes store the data and files while the computing nodes are considered homogeneous and hosts one or more VMs. The application program in the VM and the underlying operating system together are termed as workload. A server node running VMs is known as openbox and idle server node is known as closebox. In EnaCloud, the workloads are aggregated tightly to minimize the number of openboxes. The applications have varying resource demands and so workload resizing is included in it. It finds a solution for remapping workloads to the resource nodes through migration whenever a workload arrives, departs or resizes. The migration here mainly has two aims: first to reduce the number of openboxes and second is to minimize the migration time [4].

C. Application Performance Management in Virtualized Server Environments

Gunjan et al., in [6], [23] proposed Dynamic Management Algorithm (DMA) that is based on Polynomial-Time Approximation Scheme (PTAS) heuristic algorithm. The algorithm operates by maintaining two types of ordering lists, which are migration cost list and residual capacity list. The PMs are sorted according to the increasing order of their residual capacities across any resource dimension like CPU. The VMs on each PM are sorted according to the increasing order of their resource utilization like CPU usage. Migration costs of the VMs are determined based on their resource usage i.e. high usage implies high costly migration. Whenever a hotspot is detected on a PM due to violation of upper threshold, VM with least resource usage is chosen for migration to target host which has the least residual capacity to house it. If a PM cannot accommodate the VM, next PM in the sorted order is checked. Similarly, if the VM cannot be accommodated by any of the candidate target PMs, next least usage VM from the sorted order is checked.
Whenever coldspots are detected, the least usage VMs across all the underloaded PMs is chosen and migrated to a targeted PM, only if addition of the new VM increases the variance of residual capacities across all the PMs, else we choose the next VM in order. If there is no residual space left for the chosen VM, then the heuristic for coldspot mitigation stops.
Variance is defined as follows:

$$\text{variance}, R(t) = \frac{(\text{mean}-\text{rescpu})^2+(\text{mean}-\text{resmem})^2+(\text{mean}-\text{resnet})^2\ldots}{(m-1)} \quad (3)$$

$$\text{mean} = \frac{\text{rescpu}+\text{resmem}+\text{resnet}+\cdots}{m} \quad (4)$$

$$r_n = \sqrt{\text{var}_{p1}^2 + \text{var}_{p2}^2 \ldots + \text{var}_{pn}^2} \quad (5)$$

In above equation 4, mean is defined as the average of normalized residual capacities across 'm' different resources like cpu, memory, networks, etc. rescpu , resmem , resnet … stands for residual capacities across different resource dimensions. In equation 5, r_n is the magnitude of the vector which comprises of the individual variances across 'n' physical machines.
Khanna's Algorithm packs the VMs as tightly as possible trying to minimize the number of PMs by maximizing the variance across all the PMs. Thus, Khanna's algorithm minimizes power consumption by detecting underutilization in the managed using Max-Min thresholds selection model. When the resource usage of a running PM violates a minimum predefined threshold value, the algorithm tries to pack the running VMs as close as possible thus trying to minimize the number of running physical machines.

D. A Load Aggregation Method

One important method for reducing the energy consumption in data centers is to consolidate the server load

within a few number of physical machines while switching off the rest of the physical systems. Usually this is achieved by using virtualization of the systems.

Daniel Versick et. al [6] proposed this load aggregation method which uses some ideas of K-means partitioning clustering algorithm that can compute the results quickly. The K-means chooses cluster centers within an n-dimensional space randomly and the distances between the cluster centers and vertices is calculated. The algorithm has three steps: Initialization, Iteration and Termination. First, the number of clusters is calculated based on resource needs. Some of the physical machines are defined as cluster centers. Each cluster center represents one cluster. A cluster consists of physical machines hosting a number of virtual machines. Virtual Machines are added to the cluster VM list of nearest cluster center that can fulfill necessary requirements. If the clusters cannot fulfill the requirements, the virtual machine are added to new cluster with still unused physical server as cluster center. Every virtual machine is assigned to a cluster center. A new cluster center for every cluster which is nearest physical machine is calculated. If the cluster centers gets changed during last iteration and if the iterations are not at its maximum, then the empty clusters are used again to add the virtual machines. Else, the VMs of a cluster are migrated to the physical machines representing a cluster center. And at last, the physical which are not cluster centers are turned off.

E. Entropy: A Consolidation Manager for Clusters

Entropy proposes a consolidation algorithm based on constraint problem solving. The main idea of the constraint programming based resource manager is to formulate the VM resource allocation problem as constraint satisfaction problem, and then applies a constraint solver to solve the optimization problem. The ability of this solver to find the global optimum solution is the main motivation to take this approach. Entropy resource manager utilizes Choco constraint solver to achieve the objectives of minimizing the number of the running nodes and minimizing the migration cost. Entropy iteratively checks optimality constraint i.e. the current placement uses minimum number of the running nodes. If Entropy is successful in constructing a new optimal placement (uses fewer nodes) at VM packing problem (VMPP) phase, it will activate the re-allocation. Entropy employs a migration cost model that relates memory and CPU usage with migration context. High parallelism migration steps increases the cost. Using constraint programming techniques facilitates the task of capturing such context in two phases. In the first phase, Entropy computes a tentative placement (mapping of VMs to PMs) based on the current topology and resource usage of PMs and VMs and reconfiguration plan needed to achieve the placement using minimum number of PMs required. In the second phase, it tries to improve the reconfiguration plan by reducing the number of migrations required. Since obtaining the placement and reconfiguration may take a considerable amount of time, the time given to the CSP solver is defined by the users, exceeding which whatever immediate value the solver has computed is considered for dynamic placement of VMs. VMs are classified as active or inactive based on their usage of CPU with respect to thresholds set. The author define a viable configuration as one in which every active VM present in the cluster has access to sufficient cpu and memory resources on any PM. There can be any number of inactive VM on the PM satisfying the constraint. The CSP solver takes this viable condition into account in addition to the resource constraints, while procuring the final placement plan. However, considering only viable processing nodes and Cpu-Memory Resource model is the limitation of the Entropy model [7], [23].

F. Adaptive Threshold-based Approach

Anton Beloglazov and R. Buyya [2] proposed a novel technique for dynamic VM Consolidation based on adaptive utilization thresholds which ensures a high level of service level agreements (SLA) providing high quality service to customers and dealing with the energy-performance trade-offs. Fixed values of thresholds are not suitable in environments when dealing with dynamic and unpredictable workloads. The system should be able to automatically adjust in such a way as to handle the workload pattern exhibited by the application. So, there is need for auto-adjustment of the utilization thresholds based on the statistical analysis of the historical data collected during the VM's life time. The CPU utilization of a host is the summation of the utilizations of the VMs allocated to that host and can be modelled by t-distribution. The data of each VM's CPU utilization is collected separately. This, along with the inverse cumulative probability function for the t-distribution enable to set the interval of the CPU utilization. The lower threshold is the same for all the hosts. The complexity of the algorithm is proportional to the sum of the number of non-over-utilized hosts plus the product of the number of over-utilized hosts and the VMs allocated to the over-utilized hosts.

G. Sercon: Server Consolidation Algorithm using live migration of virtual machines for Green Computing

This server consolidation algorithm aimed at consolidating the virtual machines (VMs) so that minimum nodes (physical machines) are used and reducing the number of VM migrations. Certain constraints are taken into account in Sercon. They include compatible virtualization software, comparable CPU's types, and similar n/w topology and shared storage usage on both source and destination nodes, choosing the right value of CPU threshold to prevent performance degradation and migration of VM is done if it results in releasing a node. The algorithm goes like this: The nodes based on loads by VMs are sorted in decreasing order. Then, the VMs on the least loaded node in this list are selected as a candidate for migration and are again sorted according to their weights. They are allocated to the most loaded node first and so on, thus trying to compact them and so can release the least loaded node. By using this method, we can avoid numerous migrations which might otherwise be necessary if there are nodes that are still least loaded. These steps are repeated until no more migrations are possible. CPU and Memory are considered for representing load in VMs [3], [23].

Table I: Energy Efficient Server Consolidation Heuristics

Algorithm	Goal	Platform Used	Resource Considered
Sandpiper [25]	Hotspot Mitigation	Xen	CPU, memory & network
EnaCloud [4]	Minimize energy consumption, Application	Xen VMM	Memory, Storage

	scheduling		
Khanna's Algorithm [6]	Server Consolidation	VMware ESX	CPU, Memory
A Load Aggregation Method [6]	Minimize energy consumption, minimize running physical machines	Not yet implemented	Memory, Storage
Entropy [7]	Server Consolidation	Xen 3.0.3	CPU, Memory
Adaptive Threshold Based Approach [2]	Dynamic Consolidation of VMs with minimum SLA violations, no. of migrations	Cloudsim	CPU
Sercon [3]	Server Consolidation	Sercon Software	CPU, Memory
Energy Efficient VM consolidation for Cloud Computing [19]	Energy efficient storage migration, live migration of VMs	Eucalyptus	Storage
MiyakoDori [22]	Server Consolidation	Qemu/KVM	Memory
Memory Buddies [26]	Server Consolidation & Hotspot Mitigation	VMware ESX	Memory

H. Energy Efficient VM Consolidation in Eucalyptus

Pablo Graubner et. al [19] discussed this energy efficient VM Consolidation in Eucalyptus and is based on the storage synchronization. VM live migration and storage synchronization phases are introduced explicitly instead of permanent synchronization of the disk image via network. This technique leverages the concept of Distributed Replicated Block Device (DBRD), which is typically used for high availability data storage in a distributed system. The DBRD modules works in two modes: Stand-alone and Synchronized. In stand-alone mode, all the disk accesses are passed to the underlying disk driver, while in synchronized mode, disk writes are both passed to underlying disk driver and sent to a backup machine through a network while the disk reads are served locally. A multilayered root file system (MLRFS) is used for the virtual machine's root image. The basic image is cached on a local disk. The local modifications are stored on a separate layer and then they are overlaid with the basic image transparently and form a single logical file system using Copy-on-Write (COW) mechanism. Thus, only the local modifications are sent during the disk synchronization phase.

I. MiyakoDori: A Memory Reusing Mechanism for Dynamic VM Consolidation

MiyakoDori proposes a memory reusing mechanism to reduce the amount of transferred data in a live migration process. In the case of dynamic VM consolidation, a VM may migrate back to the host where it was once executed. The memory image of the VM is left on the host when it migrates out from the host and the image will be reused when the VM migrate back to host later. The few amounts of data contribute to a shorter migration time and greater optimization by VM placement algorithms. Evaluations showed MiyakoDori reduces the amount of transferred memory and total migration time of a live migration and thus reduces the energy consumption of a dynamic VM consolidation system [22], [23].

J. Memory Buddies: Exploiting Page Sharing for Smart Colocation in Virtualized Data Centers

Memory Buddies is a memory sharing aware placement system for virtual machines. It is a memory fingerprinting system to efficiently determine the page sharing potential among a set of VMs, and compute more efficient placements. It makes use of live migration to optimize VM placement as workload changes. Memory buddies detects sharing potential to realize these benefits. The memory buddies system consists of a nucleus, which runs on each server, and a control pane, which runs on distinguished control server. Each nucleus generates a memory footprint of all memory pages within the VMs resident on that server. This fingerprint represents the page-level memory contents of a VM in a way which allows efficient calculation of the number of pages with identical content across two VMs. The control plane is responsible for virtual migration placement and hotspot mitigation. For placing a virtual machine, it compares the fingerprint of that VM against server fingerprints in order to determine a location for it which will maximize sharing opportunities. It then places the VM on that server, initiating migrations if necessary. The control plane interacts with the VMs through a VM management API such as VMware's Virtual Infrastructure or the libvirt API.

Memory Buddies server consolidation algorithm opportunistically identifies servers that are candidates for shutting down and attempts to migrate virtual machines to hosts with high sharing opportunities. In doing so, it attempts to pack VMs onto servers so as to reduce aggregate memory footprint and maximize the number of VMs that can be housed in the data center. Once the migrations are completed the consolidation candidates can be retired from service or powered down until new server capacity is required, thereby saving on operational costs. The consolidation algorithm runs periodically to check the list of hosts which are candidates for consolidating if its mean usage remains below a low threshold for an extended duration. The system considers only memory usages when identifying consolidation candidates. Once the consolidation candidates are identified, the algorithm determines a new physical server to house each VM. To do so, VMs are arranged according to their decreasing order of the memory sizes and consider them one at a time. Firstly for each VM, the algorithm determines the set of feasible servers in the data centers and then the host which will provide the greatest level of sharing is selected for each VM. Once new targets have been determined for each VM on the consolidation servers, actual migration is performed using live migration. Live migration ensures transparency and near-zero down-times for the application executing inside the migrated VMs. To ensure minimum impact of network copying triggered by each migration of application performance, the algorithm places a limit on the number of the concurrent migrations; once each migration

completes, a pending one is triggered until all VMs have migrated to their new hosts. The servers are then powered off and retired or moved to a shutdown so that they can be reinitialized later if memory requirements increase [23], [26].

III. CONCLUSION

Energy Management is the one of the most challenging tasks faced by the infrastructure providers in cloud data centers. Due to subsequent increase in the data and processing, it is not possible to keep control over the energy consumption as the performance will be primarily affected. Some of the existing heuristics for energy efficient VM live migration were studied. All these heuristics mainly focus on reducing the energy consumption. Overcoming all the barriers for energy consumption is not possible as each of the heuristics through light on different parameters, though with certain disadvantage of their own.

More Heuristics can be proposed to reduce the energy consumption and to overcome the energy-performance trade-offs.

REFERENCES

[1] Anju Mohan and Shine S, "Survey on Live VM Migration Techniques", *International Journal of Advanced Research in Computer Engineering & Technology*, vol 2, Jan. 2013.

[2] Anton Belaglozavet. al, "Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers", *in Proceedings of the 8th International Workshop on Middleware for Grids, Clouds and e-Science, ACM 2010*.

[3] Aziz Murtazaer and Sangyoon Oh, "Sercon: Server Consolidation Algorithm using live migration of virtual machines for Green Computing", *IETE TECHNICAL REVIEW*, vol 28, May-Jun. 2011.

[4] Bo Li, Jianxin Li, JinpengHuai, TianyuWo, Qin Li, Liang Zhong, "EnaCloud: An Energy Saving Application Live Placement Approach for Cloud Computing Environments", *International Conference on Cloud Computing IEEE 2009*.

[5] C. Clark, K. Fraser, S. Hand, J. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfiled, "Live Migration of Virtual Machines", *In Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation*, vol. 2, pp. 286, 2005.

[6] Daniel Versick, DjamshidTavangarian, "Reducing Energy Consumption by Load Aggregation with an Optimized Dynamic Live Migration of Virtual Machines", *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, IEEE 2010*.

[7] FabeinHermenier, Xavier Lorca, Jean-Marc Menuad, Gilles Muller and Julia Lawall, "Entropy: A Consolidation Machine Manager for Clusters", *in Proceedings of the 2009 ACM SIGPLAN/SIGOPS Internal Conference on Virtual Execution Networks, VEE, 2008*, pp.41-50, 2009.

[8] Gartner report, financial times, 2007.

[9] Gunjan Khanna, Kirk Beaty, GautamKar and AndrzejKochut, "Application Performance Management in Virtualized Server Environments", *10th IEEE/IFIP Conference on Network Operations and Management in Virtualized Server Environments, NOMS*, 2006.

[10] http://searchstorage.techtarget.com.au/articles/28102-Predictions-2-9-Symantec-s-Craig-Scroggie

[11] J. Kaplan, W. Forrest, and N. Kindler, "Revolutionizing data center energy efficiency", McKinsey Company Tech Report, pp.15, July 2008.

[12] Jonathan Koomey, "Growth in data center electricity use 2005 to 2010", Analytics Press, Tech. Report, 2011.

[13] JyothiSekhar, GetziJeba and S. Durga, "A survey on Energy Efficient Server Consolidation through VM Live Migration", *International Journal of Advances in Engineering and Technology*, vol 5, pp.515-525, Nov. 2012.

[14] Kejiang Ye, Xiaohong Jiang, Dawei Huang, Jianhai Chen, and Bei Wang, "Live migration of Multiple Virtual Machines with Resource Reservation in Cloud Computing Environments" , *in Proceedings of 2011 IEEE 4th International Conference On Cloud Computing*, pp. 267-274, 2011.

[15] L. Kleinrock. A Vision for the Internet. ST Journal of Research, 2(1):4-5, Nov. 2005.

[16] M. Nelson, B. Lim, and G. Hutchins, "Fast transparent migration for virtual machines", i*n Proceedings of the Annual Conference on USENIX Annual Technical Conference*, pp. 25, 2005.

[17] Marios D. Dikaiakos, George Pallis, DimitriosKatsaros, PankajMehra, Athena Vakali, "Cloud Computing: Distributed Internet Computing for IT and Scientific", *IEEE Internet Computing*, Vol. 13, No.5, pp.10-13, Sept.2009.

[18] NIST Definition of Cloud Computing v15, www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf Retrieved 14 Oct, 2012.

[19] Pablo Graubner, Matthias Schmidt, Bernd Freisleben, "Energy-efficient Management of Virtual Machines in Eucalyptus", *4th International Conference on Cloud Computing, IEEE 2011*.

[20] P. Barhamet. al, "Xen and the Art of Virtualization", in Proc. of 19th ACM Symposium on Operating Systems Principles (SOSP 2003), Bolton Landing, USA.

[21] Rongbo Zhu, Zhili Sun, Jiankun Hu, "Special Section: Green Computing", *Future Generation Computer Systems*, pp.368-370, 2012.

[22] SoramichiAkiyama,TakahiroHirofuchi,Ryousei Takano, Shinichi Honiden, "MiyakoDori: A Memory Reusing Mechanism for Dynamic VM Consolidation", *Fifth International Conference on Cloud Computing,IEEE 2012*.

[23] Susheel Thakur, Arvind Kalia and Jawahar Thakur, "Server Consolidation Algorithms for Cloud Computing Environment: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, vol 3(9), September 2013.

[24] TharamDillion, Chen Wu, Elizabeth Chang, "Cloud Computing: Issues and Challenges", 24th IEEE International Conference on Advanced Information Networking and Applications, IEEE Computer Society 2010.

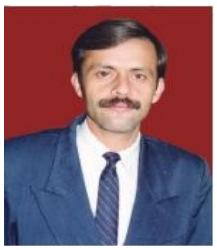
[25] Timothy Wood, PrashantShenoy, ArunVenkataramani and MazinYousif, "Sandpiper: Black-box and Gray-box Resource Management for Virtual Machines", *Journal of Computer Networks*, vol.53, pp.2923-2938, Dec.2009.

[26] T.Wood, G. Tarasuk-Levin, PrashantShenoy, Peter desnoyers, Emmanuel Cecchet and M.D.Corner "Memory Buddies:Exploiting Page sharing for Smart colocation in Virtualized Data Centers" i*n Proc. of the ACM SIGPLAN/SIGOPS International conference on Virtual execution environments,VEE*,pages 31–40, New York, NY, USA, 2009.

[27] V. Sarathy, P. Narayan, and Rao Mikkilineni, "Next Generation Cloud Computing Architecture- Enabling Real-time Dynamism for Shared Distributed Physical Infrastructure", *19th IEEE International Workshops on enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE*, pp.48- 53, June 2010.

[28] World Energy Outlook 2009 FACT SHEET. http://www.iea.org/weo/docs/weo2009/factsheetsWEO2009.pdf

Susheel Thakur received his B.Tech (I.T) Degree from University Institute of Information Technology, Himachal Pradesh University, Shimla, India and pursuing M.Tech (CS) Degree from Department of Computer Science, Himachal Pradesh University, Shimla, India. His field of Interest are Cloud Computing, Database Management System and Software Engineering.



Arvind Kalia is currently working as Professor in Department of Computer Science, Himachal Pradesh University, Shimla (H.P) India. He is having 25 years of teaching and research experience. His area of interest includes Software Engineering, Computer Networks, Data mining and Cloud Computing. He is a member of different science bodies like CSI, Indian Science Congress, etc.

Jawahar Thakur is currently working as Associate Professor in Department of Computer Science, Himachal Pradesh University, Shimla, India. He has received his B.Tech (CSE) Degree from NIT Hamirpur, India and M.Tech (CSE) from NITTTR, Panjab University Chandigarh. He is currently pursuing PhD in Computer Networking.

