Comparison of SLA based Energy Efficient Dynamic Virtual Machine Consolidation Algorithms

Heena Kaushar
Student
Bhopal Institute of Technology & Science,
Bhopal

Pankaj Ricchariya
Professor ,Head & Guide
Bhopal Institute of Technology & Science
Bhopal

Anand Motwani
Assistant Professor & Head
NRI Institute Of Research & Technology,
Bhopal

ABSTRACT
Cloud computing emerged as need for rapidly increasing computational power thus results in greater power consumption, increased operational costs and high carbon footprints to environment. A key issue for Cloud Providers is to maximize their profits by minimizing power consumption along with SLA considerations of hosted applications. Dynamic Virtual Machine (VM) consolidation is promising approach for reducing energy consumption by dynamically adjusting the number of active machines to match resource demands but it is one of the most important challenges in the cloud based distributed systems. In this work, the researchers tried to investigate “SLA and Energy-Efficient Dynamic Virtual Machine (VM) Consolidation” that meets Quality of Service expectations and Service Level Agreements (SLA) requirements. The analysis of VM consolidation algorithms based on various heuristics on legitimate host is presented as key contribution of this work. We also present a comparative analysis and results by conducting a performance evaluation study of various existing energy efficient VM consolidation techniques using real world workload traces from more than a thousand VMs using CloudSim toolkit. This paper is aimed at helping cloud providers analyze several power characteristics of their own technologies as well as pre-existing IT resources to identify their favorability in the migration to the new energy efficient cloud architectures. The results also helps in analyzing the existing frameworks and offers substantial energy savings while effectively dealing with firm QoS requirements negotiated by SLA.

General Terms
Cloud Computing, Cloud Provider, Energy Efficiency, SLA, Virtual Machine Consolidation,

Keywords
Cloud Computing (CC), Cloud Providers, Energy, Energy efficient, Quality of Service (QoS), Service Level Agreements (SLA), Virtual Machine (VM), VM Consolidation

1. INTRODUCTION
Today data centers consume tremendous amount of energy in e-work, e-commerce and e-learning in terms of power distribution and cooling. As every computing resource at Data Center produces heat, also for each watt of power consumed by computing resources, an additional 0.5-1W is required for the cooling system [1]. But it is true that, Information and communications technology (ICT)-related development can’t move forward unless the energy to power the technology is sufficient, affordable, and reliable, and the ICT equipment is engineered to serve this market and the picture is drab for datacenters that consumes high-energy [2]. In particular, cloud data centers often comprise different machines with different capacities, capabilities and energy consumption characteristics. The workloads running in these data centers typically consist of diverse applications with different objectives and resource requirements. Meanwhile, a limited budget makes providers create efficient cloud systems that utilize the computational powers of the clouds while minimizing their energy consumptions and environmental footprints.

Currently, resource allocation in a Cloud data center aims to provide high performance while meeting SLAs, without focusing on allocating VMs to minimize energy consumption [3]. But the Energy-Efficient operation of servers at data centers is achieved in number of ways such as improvement of applications’ algorithms, energy efficient hardware etc. One of the current key technologies contributing in energy efficiency and increasing ‘Return on Investment’ (ROI) includes virtualization technology [4]. Dynamic Consolidation of VMs is a promising approach for reducing energy consumption by dynamically adjusting the number of active machines to match resource demands. It includes algorithm for determination of overloaded and under-loaded hosts on basis of dynamic heuristics followed by the selection and placement of the VMs on legitimate host. The reduction in energy usage is achieved by turning off (or hibernating) unused nodes after dynamic consolidation of VMs. The objective of the VM reallocation is to minimize the number of physical nodes (i.e. the amount hardware in use) serving current workload to save energy [5]. Generally VMs moved to those locations, where they can operate in the most energy-efficient way. However, despite extensive studies of the problem, existing solutions for dynamic consolidation have not fully considered the heterogeneity of both workload and machine hardware found at cloud data centers.

In the case of a Cloud as a commercial offering to enable crucial business operations of companies, there are critical Quality of Service (QoS) parameters to consider in a service request, such as time, cost, reliability and security [6]. CC must ensures Quality of Service (QoS) requirements negotiated through Service Level Agreements (SLAs) agreed between interacting entities i.e. cloud providers, consumers and brokers. SLAs specify the agreements on the QoS, such as deadline constraints between said entities [7]. Furthermore, the cloud computing has opened up a cutting edge of challenges by introducing a different type of resource provisioning techniques. Although the optimization of VM consolidation and provisioning policies offer improvements like increasing cloud provider's profit, energy reserves and load balancing in large datacenters, the consolidation has to be done reasonably to achieve energy efficiency and performance.

The remainder of the paper is organized as follows. Section 2 explains the VM Consolidation. Section 3 discusses techniques for dynamic consolidation of VMs based on various heuristics (Related work), along with some areas of energy-efficiency research based on a cloud computing perspective. System and Power model used in this work is introduced briefly in Section 4. A performance analysis and comparison of some "SLA and Energy Aware Virtual Machine Consolidation" techniques is meticulously presented in Section 5. Finally Section 6 concludes the paper giving future directions.

2. VIRTUAL MACHINE CONSOLIDATION

The Dynamic VM consolidation framework consists of 3 processes: (1) host overload and under-load detection; (2) VM selection; and (3) VM placement. In VM consolidation process, operators consolidate all the VMs into a smaller number of physical servers, aiming at minimizing the total required servers with the constraints imposed by SLA, server capacity etc. VM consolidation in Compute Clouds has been tackled in both commercial products and research work [8] as dynamic consolidation of virtual machines (VMs) is an effective way to improve the utilization of resources and energy efficiency in cloud data centers [5].

2.1 Host Overload and Under-Load Detection

As entire description of Host Overload and Under-Load Detection methods is out of scope of this paper, we mention here names of few methods only. Following are few Adaptive Utilization Threshold based methods to determine over-utilized and under-utilized hosts [9]:

(i) Median Absolute Deviation

(ii) Interquartile Range

(iii) Local Regression.

2.2 VM Selection and Placement

The next step after the determination of overloaded and under-loaded host is determination of the particular VMs to migrate from that host. This problem of selection is solved by iteratively applying VM selection algorithms until the host is considered as not being overloaded. As entire description of VM Selection and Placement is out of scope of this paper, we mention here names of few policies. Following are few VM Selection policies adapted from [10, 3, 9] are:

(i) The Minimum Migration Time Policy.

(ii) The Maximum Correlation Policy

(iii) The Random Choice Policy

(iv) Highest Potential Growth (HPG)

The last step is to find a new placement of the VMs selected for migration from the overloaded and under-loaded hosts. For VM placement it is reasonable to apply a heuristic, such as the Best Fit Decreasing (BFD) algorithm [3], which has been shown to use no more than (11/9 .OPT + 1) bins (where OPT is the number of bins provided by the optimal solution) [11]. The destination hosts is chosen for placement of selected VMs for migration, in order to minimize power consumption.

The methods to determine over-utilized and under-utilized hosts, and policies to select a VM to be migrated, can be combined to form various strategies. In this work various Virtual machine consolidation techniques based on various heuristic for deciding the time to migrate VMs from a host, in cloud data center, are surveyed in different aspects.

3. RELATED WORK

Early work in energy efficient resource management is usually addressed to wireless devices with the objective of improving battery lifetime. A description of different techniques and approaches for reducing energy consumption is surveyed in [12]. With an enormous growth of virtual computing environments such as Clouds the context has been shifted to data centers. A large number of virtual machines (VMs) for diverse application workload requests are created at data center. Each VM is provisioned with certain amount of computing resources like storage, memory, bandwidth etc. proportionate with workload requirements. But, it becomes hard for the traditional schemes to make a deterministic estimate of resource demands.

Anton Beloglazov et al. [13] presented a survey of research in energy-efficient computing. The architectural principles for energy-efficient management of Clouds; energy-efficient resource allocation policies and scheduling algorithms considering QoS expectations and power usage characteristics of the devices; and a number of open research challenges, are addressed. This works substantially contribute to both resource providers and consumers. The approach is validated by conducting a performance evaluation study using the CloudSim [9, 10] toolkit showing significant cost savings and demonstrates high potential for the improvement of energy efficiency under dynamic workload scenarios.

Beloglazov and Buyya [4] have proposed a novel technique for dynamic consolidation of VMs based on automatically adjusting (adaptive) utilization thresholds, which ensures a high level of meeting the Service Level Agreements (SLA). The SLA violation is less than 1% and good results achieved in number of VM migrations and energy consumption. The efficiency validation of the proposed technique has been done with different kinds of workloads using workload traces from more than a thousand PlanetLab servers.

In another work Beloglazov and Buyya [5] have presented a novel technique for dynamic consolidation of VMs which ensures Service Level Agreements (SLA) and based on adaptive utilization thresholds. The propose approach based on a Markov chain model that optimally solves the problem of host overload detection under the specified QoS goal, for any known stationary workload and a given state configuration. The algorithm is heuristically adapted to handle non-stationary workloads. The extensive work has been simulated on more than a thousand VMs.

Beloglazov et al. [9] presents a heuristics for dynamic consolidation of VMs based on an analysis of historical data from the resource usage by VMs. To calculate the upper CPU utilization threshold a statistical method is used. These statistical methods to determine over-utilized and under-utilized hosts, and policies to select a VM to be migrated, can be combined to form various strategies. The destination hosts is chosen in order to minimize power consumption.

The authors [16] applied weighted linear regression to predict the future workload and proactively optimize the resource allocation to implement an energy-aware dynamic VM consolidation framework. The system mainly focused on web applications. The SLAs are defined in terms of the response time. A lot of VM consolidation techniques based on an analysis of historical data from the resource usage by VMs. Bobroff et al. [17] proposed a forecasting technique to determine server overload. The forecasting is based on time-series analysis of historical data.

Yuxiang Shi et al. [18] have achieved the energy efficiency objective by dynamically allocating resources based on utilization analysis from resource utilization log and prediction methods known as "Linear Predicting Method" (LPM) and "Flat Period Reservation-Reduced Method" (FPRRM). The M/M/1 queuing theory is used for predicting method and has better response time and less energy-consuming. Experimental evaluation performed on CloudSim [9, 10] simulator to demonstrate the proposed methods. This approach does not consider diverse workloads that are created at data centers as it is also based on previous utilization analysis.

Verma et al. [19] stated that due to unknown types of applications running on VMs it is not possible to build the exact model of a mixed workload. So the authors represented the utilization of the CPU by a VM as a uniformly distributed random variable. They formulated the problem of power-aware dynamic placement of applications in virtualized heterogeneous systems as continuous optimization. The VM placement is optimized to minimize the power consumption and maximize the performance at each time frame

4. SYSTEM AND POWER MODEL

The targeted system for comparison purpose considered in this work is an IaaS environment represented by a large-scale data center consisting of M heterogeneous physical nodes as presented in [14, 20] (see Figure 1). The global manager resides on the master node has two tasks (i) to collect statistics from the local managers to maintain system's resource utilization and (ii) to issue VM migration commands to optimize the VM placement. Virtual Machine Manager (VMM) performs actual migration of VMs and takes decisions to alter the power modes of the nodes. As an idle server consumes approximately 70% of the power consumed when it is fully utilized. We used the power model and energy consumption model as described in [4, 3], which defines the power consumption as a function of the CPU utilization P(u) as shown in (1) and total energy consumption by a server as defined in (2), where Pmax is the maximum power consumed; k is the fraction of power consumed by an idle server; and u is the CPU utilization.

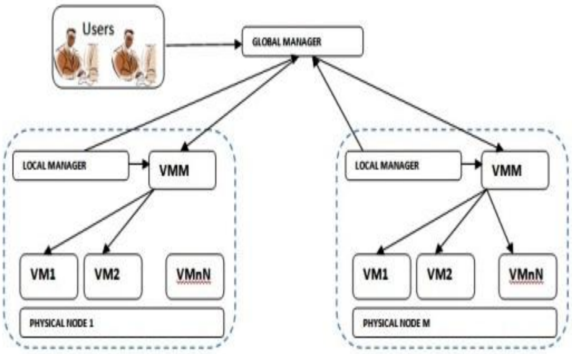

Fig 1: System Model

$$P(u) = k.P_{max}+(1 - k).P_{max}.u = P_{max}.(0.7+0.3.u) \quad (1)$$

$$E = \int_t P(u(t))dt \quad (2)$$

5. COMPARISON AND PERFORMANCE ANALYSIS

It is extremely difficult to conduct exhaustive large-scale experiments on a real infrastructure which is required to evaluate and compare the algorithms discussed in this work. So, simulations have been done on CloudSim [14, 15] toolkit which is a modern simulation framework aimed at Cloud computing environments. Here in this paper comparative study of following combinations of over and under-utilized host determination method, and VM selection is done. Also, we compared all the combinations with benchmark policy Dynamic Voltage and Frequency Scaling (DVFS).

In section 5.3.1 Minimum Migration Time [9] VM selection policy is combined with three over and under-utilized host determination methods and comparison is done on basis of various parameters. Similarly, in section 5.3.2 Maximum Correlation [9] VM selection policy is combined with three over and under-utilized host determination methods and comparison is done on basis of various parameters. The conventions used for representing algorithms are shown in Table – 1.

Table – 1 Conventions used

CONSOLIDATION POLICY	CONVENTION
Maximum Correlation + Threshold Algorithm	**ThrMc**
Maximum Correlation + Median Absolute Deviation	**MadMc**
Maximum Correlation + Interquartile Range	**IqrMc**
Minimum Migration Time + Threshold Algorithm	**ThrMmt**
Minimum Migration Time + Median Absolute Deviation	**MadMmt**
Minimum Migration Time + Interquartile Range	**IqrMmt**

5.1 Experimental setup

The methods to determine over-utilized and under-utilized hosts, and policies to select a VM to be migrated, can be combined to form various strategies. In this work various Virtual machine consolidation techniques based on various heuristic for deciding the time to migrate VMs from a host in cloud data center are surveyed in different aspects.

The simulation platform chosen for experimentation purposes is CloudSim 3.0.2, as it is a modern simulation framework aimed to simulate Cloud computing environments. The simulation has been performed by defining 800 physical nodes of two types in a data center. Each node is modeled to have two CPU core with performance equivalent to 1860 Million Instructions Per Second (MIPS) each core in Type-1 and 2660 (MIPS) each core in Type-2, Random Access Memory RAM of 4 GB, 1 GB/s network bandwidth and 1 GB of storage. Instead of using an analytical model, this work utilizes real data on power consumption.

Each VM requires one CPU core with maximum of 2500, 2000, 1000, 500 MIPS, 1 GB of RAM, 1 GB/s network bandwidth and 1 GB of storage. However, during the lifetime VMs may use fewer resources creating the opportunity for dynamic consolidation. Each VM is randomly assigned a workload trace from one of the servers from the workload data. The simulations have been run on 24 hours to find out the energy consumption, SLA violation and number of VM migrations over different workload types. The utilization measurement is done at interval of 5 minutes.

5.2 Performance Metrics

5.2.1 Total energy consumption:

Total energy consumption is defined as the sum of energy consumed by the physical resources of a data center as a result of application workloads.

5.2.2 SLA violation:

When a VM cannot get the promised Quality of Service (QoS), SLA violation takes place. For example when a VM cannot get requested MIPS SLA violation issue occur.

5.2.3 Number of VM migrations:

For dynamic VM consolidation it is must to determine overloaded and under-loaded hosts and once the overloaded or under-loaded hosts found the VMs get selected for migration. The minimization of the VM migration time is more important constraint and one of the ways to achieve it is to reduce the total number of VM migrations.

5.3 Simulation Results and Analysis

5.3.1 Comparison of Dvfs, ThrMmt, MadMmt, IqrMmt

The comparison between Dvfs, ThrMmt, MadMmt and IqrMmt on basis of five parameters is shown in Table – 2. The graphs of Energy consumption, Number of VM migrations and Number of host shutdowns of each of these schemes is represented in Figure 2, 3 and 4 respectively.

Table 2. Comparison of Dvfs, ThrMmt, MadMmt, IqrMmt

Parameters	Dvfs	IqrMmt	MadMmt	ThrMmt
Energy Consumption (kWh)	803.91	188.86	184.88	191.73
No. of VM migrations	0	26476	26292	26634
Overall SLA violation (%)	0	0.07	0.08	0.07
Average SLA violation (%)	0	9.98	10.18	10.14
No. of Host shutdowns	457	5827	5759	5863

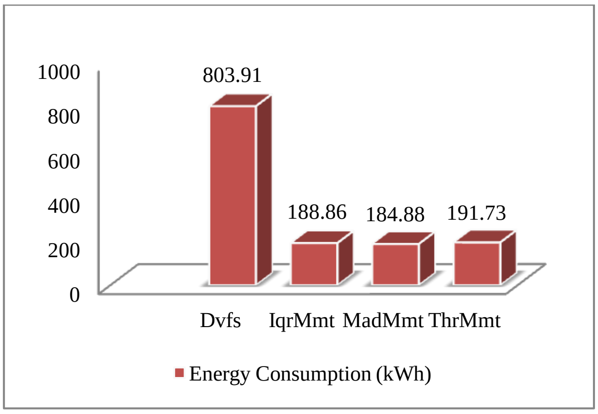

Fig 2: Energy Consumption

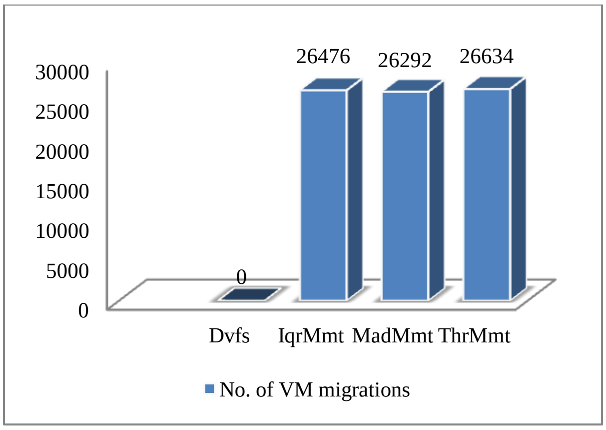

Fig 3: Number of VM Migrations

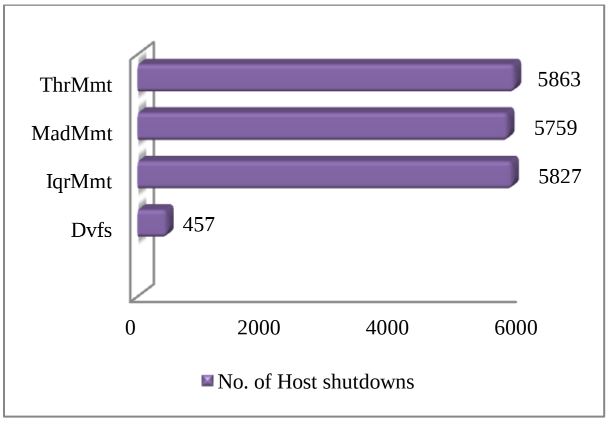

Fig 4: Number of Host Shutdowns

Numbers of Host shutdowns are very fewer in case of Dvfs but energy consumption is almost four times the said schemes.

5.3.2 Comparison of Dvfs, ThrMc, MadMc, IqrMc

The comparison between Dvfs, ThrMc, MadMc, and IqrMc on basis of five parameters is shown in Table – 3. The graphs of Energy consumption, Number of VM migrations and Number of host shutdowns of each of these schemes is represented in Figure 5, 6 and 7 respectively.

Table 3. Comparison of Dvfs, ThrMc, MadMc, IqrMc

Parameters	Dvfs	IqrMc	MadMc	ThrMc
Energy Consumption (kWh)	803.91	178.8	176.24	182.96
No. of VM migrations	0	23270	23636	24172
Overall SLA violation (%)	0	0.12	0.12	0.11
Average SLA violation (%)	0	9.83	9.76	9.94
No. of Host shutdowns	457	5428	5425	5525

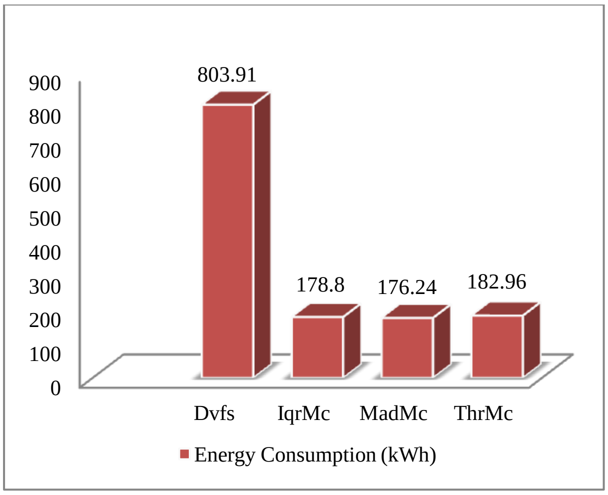

Fig 5: Energy Consumption

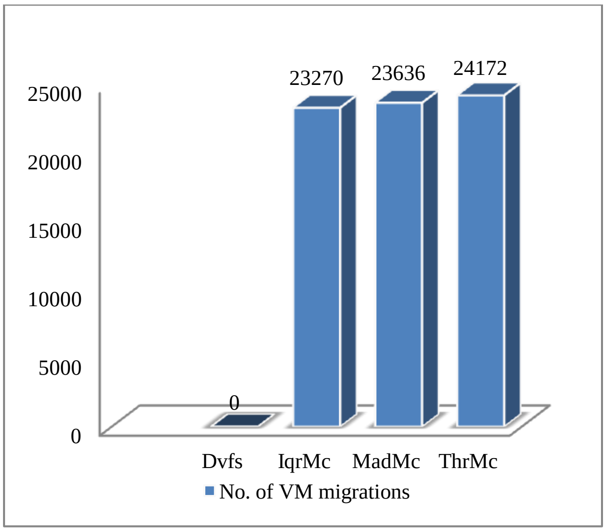

Fig 6: Number of VM Migrations

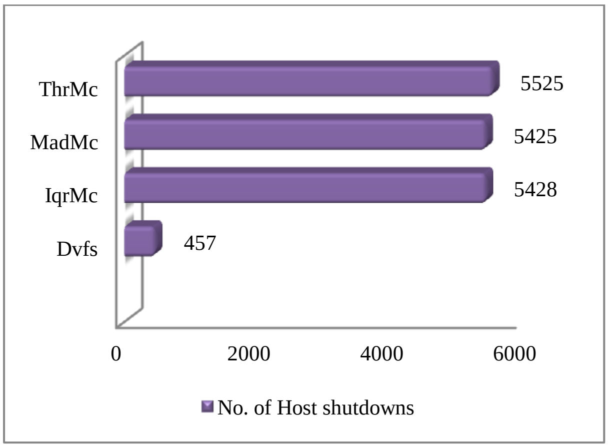

Fig 7: Number of Host Shutdowns

The Overall SLA violation in case of DVFS is 0. On the other hand Energy efficient schemes showing very little SLA violation (<1%) (See Figure 8). 'Mmt' policies along with three overload detection methods are showing up to 33% less SLA violation than 'Mc' policy.

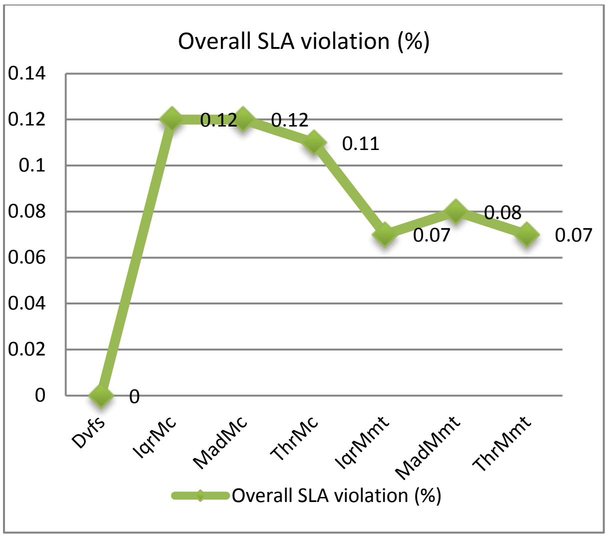

Fig 8: Overall SLA violation

The Energy consumption of all schemes shows linear pattern. 'ThrMmt' is showing highest energy consumption in all schemes discussed in this work (refer Figure 9).

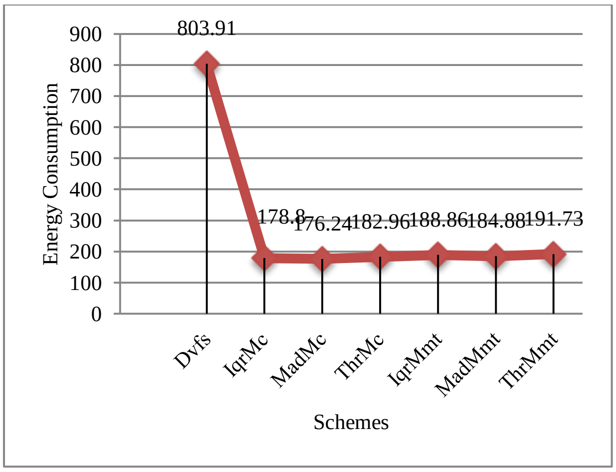

Fig 9: Comparison of Energy Consumption

6. CONCLUSION AND FUTURE WORK

This paper focuses on energy-efficient computing and present: (a) survey on "SLA and Energy-Efficient Dynamic Virtual Machine (VM) Consolidation"; (b) analysis of VM consolidation algorithms based on various heuristics; (c) comparative analysis and results by conducting a performance evaluation study of various energy efficient VM consolidation techniques using real world workload traces. The comparison is done on various parameters and with DVFS policy.

Current Cloud data centers host applications having heterogeneous requirements, from clients distributed globally and these requirements may vary over time. Cloud Providers need to provide strict QoS guarantees, which are documented in the form of SLAs. The resource consolidation techniques within a data center directly influences whether the SLAs are met. Although varied application requirements of cloud customers make scheduling and VM consolidation algorithms complex, they can be exploited to improve energy-efficiency. The focus of this work is to study energy and QoS efficient VM consolidation strategies that can be applied in a virtualized data center by a Cloud provider. In addition, it is necessary to design a solution scalable to handling thousands of users, so to evaluate the existing mechanisms, a series of simulation experiments were conducted on the CloudSim platform using real workload traces.

As a future direction, this work suggests the study of other energy efficient Resource provisioning techniques over varied workloads to make the data centers scalable and reliable in terms of QoS. Also the individual problems like host overload detection etc. of energy-efficient dynamic VM consolidation can be analyzed and modified individually. The application of optimal and near optimal approximation algorithms can be applied to workload forecasting and individual problems of Resource consolidation may prove potential. The work will encourage the researchers to perform competitive analysis of these algorithms to get theoretical performance.

7. REFERENCES

[1] Ranganathan P, Leech P, Irwin D, Chase J. Ensemble-level power management for dense blade servers. Proceedings of the 33rd International Symposium on Computer Architecture (ISCA 2006), Boston, MA, USA, 2006; 66–77.

[2] Hosman, Laura; Baikie, Bruce, "Solar-Powered Cloud Computing Datacenters," IT Professional , vol.15, no.2, pp.15-21, March-April 2013.

[3] Anton Beloglazov, Jemal Abawajy, Rajkumar Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing”, Future Generation Computer Systems Volume 28, Issue 5, May 2012, pp. 755–768, Elsevier.

[4] Anton Beloglazov and Rajkumar Buyya, “Adaptive Threshold-Based Approach for Energy-Efficient Consolidation of Virtual Machines in Cloud Data Centers”, MGC ’2010, 29 November - 3 December 2010, Bangalore, India. Copyright 2010 ACM 978-1-4503-0453-5/10/11.

[5] Anton Beloglazov and Rajkumar Buyya, “Managing Overloaded Hosts for Dynamic Consolidation of Virtual Machines in Cloud Data Centers under Quality of Service Constraints”, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 7, JULY 2013.

[6] Rajkumar Buyyaa, Chee Shin Yeoa, Srikumar Venugopala, James Broberg , Ivona Brandic, “Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility”, Future Generation Computer Systems 25 (2009) 599-616 Elsevier.

[7] Yue Gao; Yanzhi Wang; Gupta, S.K.; Pedram, M., "An energy and deadline aware resource provisioning, scheduling and optimization framework for cloud systems," Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2013 International Conference on , vol., no., pp.1,10, Sept. 29 2013-Oct. 4 2013.

[8] Meng Wang, Xiaoqiao Meng, and Li Zhang, “Consolidating Virtual Machines with Dynamic Bandwidth Demand in Data Centers”, Mini Conference IEEE INFOCOM 2011, 978-1-4244-9920-5/11/ 2011 IEEE.

[9] Anton Beloglazov, Rajkumar Buyya, “Optimal online deterministic algorithms and adaptive heuris-tics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers”, Wiley InterScience, Concurr. Comput. : Pract. Exper., 24(13):1397-1420, September 2012.

[10] Beloglazov, A.; Buyya, R., "Energy Efficient Allocation of Virtual Machines in Cloud Data Centers," Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on , vol., no., pp.577,578, 17-20 May 2010.

[11] M. Yue, “A simple proof of the inequality FFD (L)< 11/9 OPT (L)+ 1,for all l for the FFD bin-packing algorithm,” Acta Mathematicae Applicatae Sinica (English Series), vol. 7, no. 4, pp. 321–331, 1991.

[12] Hesham Hassan, Ahmed Shawky Moussa, “Power Aware Computing Survey”, International Journal of Computer Applications (0975 – 8887) Volume 90 – No.3, March 2014.

[13] Buyya R, Beloglazov A, Abawajy J. Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges” in Proceedings of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2010). Las Vegas, USA, July 2010.

[14] Buyya, R.; Ranjan, R.; Calheiros, R.N., "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities," High Performance Computing & Simulation, 2009. HPCS '09. International Conference on , vol., no., pp.1,11, 21-24 June 2009.

[15] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, Cesar A. F. De Rose, and Rajkumar Buyya, “CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms,” Software: Practice and Experience (SPE), Volume 41, Number 1, Pages: 23-50, ISSN: 0038-0644, Wiley Press, New York, USA, January, 2011.

[16] B. Guenter, N. Jain, and C. Williams, “Managing Cost, Performance, and Reliability Tradeoffs for Energy-Aware Server Provisioning,” Proc. IEEE INFOCOM, pp. 1332-1340, 2011.

[17] N. Bobroff, A. Kochut, and K. Beaty, “Dynamic Placement of Virtual Machines for Managing SLA Violations,” Proc. IFIP/ IEEE 10th Int’l Symp. Integrated Network Management (IM), pp. 119-128, 2007.

[18] Yuxiang Shi; Xiaohong Jiang; Kejiang Ye, "An Energy-Efficient Scheme for Cloud Resource Provisioning Based on CloudSim," Cluster Computing (CLUSTER), 2011 IEEE International Conference on , vol., no., pp.595,599, 26-30 Sept. 2011.

[19] Verma A, Ahuja P, Neogi A. pMapper: Power and migration cost aware application placement in virtualized systems. Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware (Middleware 2008), Springer, Leuven, Belgium, 2008; 243–264.

[20] Y. Song, “Multi-Tiered On-Demand resource scheduling for VM-Based data center” In Proc. of the 2009 9th IEEE/ACM Intl. Symp. On Cluster Computing, 155, 2009.