SLA CONSCIOUS VM MIGRATION

Cloud Computing has emerged as an alternate infrastructural, computational and service platform. It caters the users with least financial burden. An important aspect of cloud computing remains QoS, often referred as SLA. Several user jobs and applications concurrently compete for cloud resources. Cloud environment exploits virtualization for resource management. Virtualization allows cloud service providers to instantiate VM for subscribers. VMs are like producers and consumers in cloud environment. VMs consume computational, storage and network resources in cloud on behalf of jobs and tasks. Jobs of subscribers are represented as VM or VMs. In this view, VMs are producers. VMs are smallest unit of discussion in cloud environment. With limited available resources and compelling SLAs, VMs have to be migrated live during execution. Another concern in Clouds remains energy efficiency. Because of huge sized servers and communication requirements, most of the investment goes in power. This chapter presents a novel SLA aware VM Selection Scheme. Section 4.1 presents issues addressed in this chapter. Section 4.2 presents introduction to proposed scheme and justifies its inception. Section 4.3 consolidates architectural assumptions and notations. Section 4.4 describes Variance aware VM migration scheme named Var_Sel. Section 4.5 elaborates assumptions with respect to simulation. Section 4.6 discussed the performance evaluation of Var_Sel. Finally, Section 4.7 summarizes the chapter.

4.1 Issues

As green computing concerns are being raised by socio-economic groups so, energy efficiency of resources in data centers is considered at equal importance as the utilization of resources in data centers. Huge demand of energy by data centers in cloud has raised concerns about carbon imprints due to energy usage. Total cost of ownerships
has grown manifolds due to energy expenses in operation of cloud services. Migration of VMs may help cloud owners to consolidate hosts and switch-off under-utilized hosts after consolidation. A lot of research is being carried out in the field of VM migration and hosts consolidations over last few years. Usage of several approaches like greedy approach, GAs, heuristics and meta-heuristics for devising solution for VM migrations has been practiced. Mostly, centralized approach is utilized for decision making in VM migration and host consolidation. This paper proposes a novel VM selection policy based on variance statistic. The simulation of the proposal has been done exhaustively and compared the performance of scheme on several parameters of relevance. The contributions of this work are listed as follows:

1. SLA conscious application requires bandwidth consideration which otherwise is considered a granted attribute of VM.

2. Heuristic based approach coupled with SLA consciousness outperform other approaches based on resource optimization.

3. Improvement in mean execution time of VM selection from given host.

4. Host selection performance in cloud datacenter is improved as a result of load unification after VM migration.

5. Mean Time before VM migration has dropped significantly along with number of VM Migrations.

4.2 Variance based VM Migration

Recent advances in technology derived a new paradigm in parallel and distributed computing, known as cloud computing. Cloud framework behaves as service platform in the form of PaaS, IaaS and SaaS. These services are available to the users in a pay-as-you-go manner [2]. Similar services provided by technology giants like Google and Microsoft are live examples of cloud computing.
Cloud computing environment is ensemble of large scale high power data centers, huge storage and high bandwidth communication network. Datacenters comprises of sufficiently huge number of physical nodes. These physical nodes are called hosts and constitute lowest layer in cloud computing environment. Hosts are virtually shared among jobs or applications through virtualization. Virtualization illusions the availability of parallelism in computing, storage and communication resources by hypothetical machines called VMs. The allocation of multiple VMs on single physical host help consolidates the hosts in data centers. Sharing of physical hosts among different applications through virtualization offers maximum utilization, resizing of VMs, migrations of VMs to other hosts and consolidation of hosts. Live migration of VMs among hosts in data centers is newest terminology and results minimum downtime and maximum consolidation [101]. VM migration is an effort to exploit the available resources in energy efficient manners and respects consumer requirements usually specified through SLAs. Any migration policy should make best effort to minimize or eliminate SLA violations.

4.3 Architecture Assumptions and Notations

The architectural framework of cloud environment is shown in Figure 4.1. Users submit their requests in the form of task graph and user also generates the VM requests based upon the task characteristics. Each request is handled by the resource management policy that allocate appropriate amount of VMs for each workload request and place those VMs on physical servers. Scheduler interface is responsible for providing priority policies and match making services of the user request. History of all the users history will be saved in the database.

Cloud computing platform is receiving huge response from all classes of users including those of enterprise and non-enterprise users. Several thousands of requests are handled by clouds through virtual resources. This is called Virtualization. Virtualization allows us to share the resources on each server by using VMs. User’s requests are represented as one or more VMs running on hosts. VMs are the smallest unit of resource utilization. VMs are allocated or reallocated on server nodes while ensuring that VMs requirements
don’t exceed server node’s capacity. In general servers consume a lot of energy in each of its states namely, Idle and Occupied. In cloud environment, most of energy consumption is attributed to the server nodes. The energy consumption of servers can be reduced by server consolidations where server consolidation is popularly achieved through VM migrations.

Figure 4.1: Energy Efficient Scheduling Framework
Table 4.1 Notation Table

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Var_{Util}$</td>
<td>Aggregate Variance of Utilization</td>
</tr>
<tr>
<td>$Var_{Util_j}$</td>
<td>Variance of Utilization until $j^{th}$ Overloaded Host</td>
</tr>
<tr>
<td>$Var_{Util_{j+1}(j, k)}$</td>
<td>Change in Variance of Utilization after considering $k^{th}$ VM on $j + 1^{th}$ Overloaded Host</td>
</tr>
<tr>
<td>$N_{OL}$</td>
<td>Number of Over Loaded Host</td>
</tr>
<tr>
<td>$N_{i}^{VM}$</td>
<td>Specific VM on $i^{th}$ Over Loaded Host</td>
</tr>
<tr>
<td>$U_n(t)$</td>
<td>Utilization of $n^{th}$ host at instant $t$</td>
</tr>
<tr>
<td>$power_{max}^n$</td>
<td>Maximum Power consumption by $n^{th}$ host</td>
</tr>
<tr>
<td>$power^n(t)$</td>
<td>Effective Power consumption by $n^{th}$ host</td>
</tr>
<tr>
<td>$n$</td>
<td>Server Number</td>
</tr>
<tr>
<td>$N_{OL}$</td>
<td>Number of Over Loaded Host</td>
</tr>
<tr>
<td>$U_n(t)$</td>
<td>Utilization of $n^{th}$ host</td>
</tr>
</tbody>
</table>

Let $power_{max}^n$ denote the maximum power consumed by a server $n$. The idle server consumed 70% of the power consumed by the respective server when it is fully utilized. The power consumption by the server $n$ at any instant of time $t$ is [102].
\[ power_n(t) = 0.7 \times power_{\text{max}}^n + 0.3 \times power_{\text{max}}^n \times U_n \] (4.1)

Where \( U_n(t) \) represent utilization at that instant of time. Servers consume a lot of power even if they are idle. It is better if idle or lightly loaded server nodes may be vacated and switched-off. Another situation arises when a heavily loaded server halts due to over allocation of VMs resulting in SLA violations. Thus energy efficiency of clouds and reduction in SLA Violation are two main concerns in cloud computing environment. Several different approaches have been proposed which includes live VM migration and server consolidation. Considering VM Migration as one of the approach for server consolidation and reducing live server nodes, there are several related sub problems. Migration of VMs requires assessing the condition of overloading, selection of best VM and placement of selected VM on target server node. VM migration introduces its own challenges like performance degradation due to migration. Keeping these many challenges in mind, it is NP-hard problem to find the source VM and destination server efficiently.

4.4 A Novel Approach for VM Selection

VM Migration is revolutionary approach while aiming for performance improvement in multiple characteristics in cloud datacenter. Any voluminous change in performance characteristics will derive energy efficiency of data centers in clouds. Several proposals have been made recently, but most of them considered basic VM characteristics as specified in available tools. When considering VM migration, a major challenge remains to select best VM from a set of VMs allocated to a server (host) node. VM selection may be realized by applying some heuristics, meta-heuristic, evolutionary algorithm or optimization approach. The selected VM have to be migrated to another host. The selected host must be able to accommodate newly migrated VM without violating the resource limitation and threshold constraints.
4.4.1 VM Allocation Assumption

VMs are created as per user requests. All VMs on their arrival are allocated to hosts in a datacenter considering all resource constraints using different allocation policies [104] as discussed below:

Static Threshold (THR): Considers static values for both, Lower Threshold and Upper Threshold for utilization.

Median Absolute Deviation (MAD): MAD is measure of statistical dispersion. It is more robust estimator of scale than the sample variance or standard deviation of given data [104].

Inter Quartile Range (IQR): IQR also called mid-spread or middle-50 and is measure of statistical dispersion. It equals to the difference between the third and first quartile.

Local Regression (LR): Local regression is fitting simple models to localized subsets of data to build a curve that approximate the original data [104].

Local Regression Robust (LRR): To overcome the vulnerability of outliers in Least Square Method, LRR is proposed.

Migration of VM is implies a situation where server utilization has breached a specific utilization level. The usage threshold is specified the Upper Threshold and Lower Threshold of utilizations. The values of Upper Thresholds and Lower Thresholds are computed using MAD. Upper Threshold dictates a situation where no processing can take place and host experience a stand-still phase. During this time processes on the said host enters ready interrupted state. Host will restart the execution of the VMs or tasks only after load shedding on this host. Migration of a VM is most effective in this scenario. Another situation is dictated by Lower Threshold, which specifies minimum level of utilization of any host. If lower threshold is breached, it implies a best situation to shut down the host and thus migrating VMs from the said host to other hosts in the datacenter.
4.4.2 Variance Minimization based Selection (Var_Sel)

To retrieve best VM for migration, at least selection policy (scheme) is required. Most of the selection policies are based on parameter like maximization or minimization of utilization, storage, communication and their combination. A new variance minimization selection scheme, namely Var_Sel, is proposed in this work. The proposed approach is based on unifying the utilization across the hosts in a given data center and selecting a VM which reduces or maintains variance of remaining load levels of hosts. A mathematical variance based heuristics is devised for assessing various VMs on an overloaded host and select best suitable VM for migration. Figure 4.2 shows Var_Sel Algorithm. This will unify the load of hosts after migration. It would also allow host selection algorithm to converge quicker than before.

Proposed Algorithm: Var_Sel (Host-List)

1. **Begin**
2. Compute the Upper Threshold and Lower Threshold values of Utilization for the given context
3. Compute the list of Over-Loaded hosts in Over-Loaded Hosts and VM’s Utilizations on all Hosts
4. Initialize Variance-Utilization = 0, Best-VM = -1;
5. **For** each host \( h \) in Over-Loaded Hosts **do**
   **Begin**
   **If** \( (h \) is first host)
   **Begin**
   Best-VM = max \{VM-Utilization\};
   Variance-Utilization = 0;
   Add Best-VM to VM-Migration-able-List
   **Continue;**
End

Sort all VMs of each Host in decreasing order by MIPS and place in VM-List in decreasing order

For each VM in VM-List do

Begin

Let VM-Candidate be VM under consideration

Compute Current-Host-Utilization = Host-Utilization-VM-Candidate-Utilization

Compute Variance Utilization considering Current-Host-Utilization w.r.t. host 1 to h-1

If (new variance has reduced or remains unchanged)

Begin

Best-VM = VM-Candidate

Variance-Utilization is updated to new lower variance

End

End

Add Best-VM to VM-Migration-able-List

End

6. Return (VM-Migration-able-List)

Figure 4.2: Variance Minimization based Selection Algorithm (Var_Sel)

Equation {2} is used to identify VMs across host list.
\[ \text{Var}_{\text{Util}} = \min \left\{ \text{Var}_{\text{Util}_j} + \text{Var}_{\text{Util}_{j+1}}(j,k) \right\} \quad (4.2) \]

Where \( j = 1..N_{OL} \) and \( k = 1..N_{VM}^i \).

Next VM considered best for migration is that minimizes the value of \( \text{Var}_{\text{Util}} \). The algorithm in Figure 4.2 performs step corresponding to VM selection as per \{Eqn. 4.2\}. Variance based selection allows the selection of a VM which will leave overloaded hosts at almost same level of utilization. This process has to be repeated for all the VMs. To ease the scanning over all the VMs, sorted of VMs by their MIPS is considered. If we found any VM at any position in sorted order starts increasing variance then all the VMs after this position will do the same. Thus scheme need not to scan entire list of VMs. In worst case only scheme will scan over complete list. This would allow \text{Var}_\text{Sel} \text{ algorithm to conclude faster than many of VM selection algorithms which scans complete list before identification of best VM. Due to sorted list mean execution time of \text{Var}_\text{Sel} \text{ will improve in comparison to other selection schemes. Another possible gain is due to host load unification. Unified load on hosts gives an opportunity to find the host to relocate the VM faster than other selection scheme. \text{Var}_\text{Sel} \text{ will possibly affect the other functional parameters of the cloud datacenter. The simulation based analysis of the scheme is presented in next section. }

Table 4.2 Power consumption of Servers at different load levels (W)

<table>
<thead>
<tr>
<th>Server</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP ProLiant G4</td>
<td>86</td>
<td>89</td>
<td>92</td>
<td>96</td>
<td>99</td>
<td>102</td>
<td>106</td>
<td>108</td>
<td>112</td>
<td>114</td>
<td>117</td>
</tr>
<tr>
<td>HP ProLiant G5</td>
<td>93.7</td>
<td>97</td>
<td>101</td>
<td>105</td>
<td>110</td>
<td>116</td>
<td>121</td>
<td>125</td>
<td>129</td>
<td>133</td>
<td>135</td>
</tr>
</tbody>
</table>
Table 4.3 Resource Configuration assumed for VMS

<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Name</th>
<th>Processor</th>
<th>MHz</th>
<th>Cores</th>
<th>RAM</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Servers</td>
<td>HP ProLiant</td>
<td>Intel Xeon</td>
<td>1860 MHz</td>
<td>2 cores</td>
<td>4 GB</td>
<td>1GB/s</td>
</tr>
<tr>
<td></td>
<td>HP ProLiant</td>
<td>Intel Xeon</td>
<td>2660 MHz</td>
<td>2 cores</td>
<td>4 GB</td>
<td>1 GB/s</td>
</tr>
<tr>
<td>VMs</td>
<td>AmazonEC2</td>
<td>High-CPU</td>
<td>2500MIPS</td>
<td>1 core</td>
<td>1.870GB</td>
<td>1.5 MB/s</td>
</tr>
<tr>
<td></td>
<td>AmazonEC2</td>
<td>Extra Large</td>
<td>2000MIPS</td>
<td>1 core</td>
<td>1.740GB</td>
<td>.25MB/s</td>
</tr>
<tr>
<td></td>
<td>AmazonEC2</td>
<td>Small</td>
<td>1000MIPS</td>
<td>1 core</td>
<td>1.740GB</td>
<td>.75MB/s</td>
</tr>
<tr>
<td></td>
<td>AmazonEC2</td>
<td>Micro</td>
<td>500MIPS</td>
<td>1 core</td>
<td>.613GB</td>
<td>.50MB/s</td>
</tr>
</tbody>
</table>

4.5 Simulation Environment

Feasibility of the solution proposed in this paper has been established through simulation. Simulation environment is setup using CloudSim toolkit [197] configured in Eclipse on an Intel Core 2 Duo, 2.0 GHz Linux based laptop. Several assumptions were made and are described in section 4.4.1. The configurations and power consumption characteristics of the selected servers are shown in Table 4.2 [104]. Server’s and VM’s configurations are specified in Table 4.3. The simulations were carried out on real time workload instances provided by a monitoring infrastructure for PlanetLab [198].

Simulation considered overloaded detection algorithms available in CloudSim Toolkit 3.0 namely MAD, IQR, THR, LR and LRR. Simulation was carried out using 800 hosts and 1052 VMs. For comparison, Maximum Correlation, Minimum Migration Time, Maximum Utilization and Random Selection Policies were considered as the benchmarks. Some specific changes were made with respect to VM configuration that suits to SLA conscious application. Such application requires that VM be prone to
migrate under overload and under-load condition. VM migrations will help to consolidate the overloaded and under-loaded servers. VM types are based on bandwidth allocated to VMs rather than their RAM requirements.

4.6 Analysis and Discussion

Analysis of the simulation revolves around the all-round improvements experienced in various parameters of importance. The impact of VM selection and Host Load unification after VM migration will affect several direct and in-direct parameters during simulations. Majority of direct parameters include Energy Efficiency, SLA Violations, SLA Time per Active Host (SLATAH), Number of VM Migrations, and performance degradation due to VM Migrations. Besides these, indirect parameters of importance considered are Mean-execution time for Selection Algorithm, Standard Deviation of VM Selection execution time, Standard Deviation of Host Selection Execution Time, Mean Time before VM Migration and Standard Deviation of Time before VM Migration.

**Energy Efficiency**: Energy efficiency is defined as the scenario of achievement or completion of given task(s) at hand using minimum amount of energy consumption. Due to rising environmental concerns and green house effects it is important to reduce carbon footprints. Selective shutdown of underutilized hosts and maintaining QoS is important parameter and trade off as well. Figure 4.3 displays energy consumption of various selection schemes based upon Maximum Correlation (Max_Corr), Minimum Migration Time (Min_Mig_Time), Maximum Utilization (Max_Util), Random Selection (Rand_Sel) and Variance Minimization based Selection (Var_Sel) respectively. Data in table aids in clarifying the numerical magnitude of performance gain of Var_Sel against other selection schemes. Across all allocation schemes, Var_Sel has performed better than other selection schemes. The primary reason behind the improvement is attributable to less time spent in migration of VMs after selection and improved convergence time of VM selection algorithm in Var_Sel due to sorted order consideration of VMs while scanning VMs of an overloaded host.
Number of VM Migrations: Number of VM Migrations is the measure of effectiveness of VM selection scheme. Var_Sel selection algorithm is based on reducing the variance of the load of hosts after VM migration. VM Migration leaves overloaded host at almost common load level after migration. The unified load level allows allocation algorithms to find a host quicker than the case with non-uniform load levels. Moreover after initial migrations most hosts remains occupied in given threshold range. Var_Sel performs average in this evaluation of VM Migrations. Lower numbers of migrations are due to better VM selection for Migration at first place. How many phases of migrations were performed during is left as future works. Hopefully Var_Sel has reduced the number of Migration Phases. Reduced VM Migrations affected Energy consumptions. Figure 4.4 shows the performance of Var_Sel against other selection schemes.

![Energy Characteristics](image)

Figure 4.3 Energy Consumption in various VM Selection Policies
Figure 4.4: Number of VM Migrations in various VM Selection Policies

Figure 4.5: SLA Time per Active Host in various VM Selection Policies
**SLA Violation (SLAV):** SLAV is a composite parameter and is defined as the product of SLA Violation Time per Active Host (SLATAH) and Performance Degradation due to Migration (PDM). SLATAH is percentage of time, when active hosts have experienced the CPU utilization of 100%. During Migration also considerable percentage of MIPS (app. 10%) is used and results in reduction of available MIPS for use by VMs. The unified metric SLA Violation (SLAV) is product of SLATAH and PDM. Graph in Figure 4.5 presents SLATAH characteristics. Var_Sel has reduced this time to almost half than that of worst performer selection scheme. The consistent performance is apparent across all the strategies for VM allocations. SLATAH is directly dependent upon the time spent during VM migration. Due to bandwidth based classification of VMs, VM Migration is faster. Lower SLATAH will affect SLAV instances.

![Performance Degradation Characteristics](image)

**Figure 4.6:** Performance Degradation due to Migration in various VM Selection Policies
Figure 4.7: SLA Violation in various VM Selection Policies

Figure in 4.7 displays SLAV characteristics of various selection schemes under different allocation scheme. SLAV performance of Var_Sel is better than any other selection scheme. Lower values of SLA Violations are due to lower SLATAH and lowering of performance degradation due to VM Migrations. VM Migrations are fact of cloud environment. Due to reduced time in migrations and lower SLATAH, SLA violations are reduced drastically. Unifying utilization across all hosts in a data center after VM Migrations is leading cause that affected these three parameters of importance. By unified utilization achievement, number of migrations is reduced in Var_Sel. Reduced no. of Migrations reduces MIPS spent in migrations. This ultimately reduces the values in PDM. SLATAH is lower in Var_Sel due to the fact that VM migrations were performed faster than trivial cases. Lower values of both SLATAH and PDM produced combined metric SLAV. SLATAH and PDM have been drawn in {Figure 4.5 and Figure 4.6}, respectively.
Figure 4.8: Execution Time Characteristics of VM Selection Policies

Diagram in (Figure 4.8) displays execution time characteristics of Var_Sel and Max_Corr selection algorithms schemes under different allocations schemes.
Figure 4.10: Mean Time before VM Migration in various VM Selection Policies

Figure 4.11: Standard Deviation of time before VM Migration in various VM Selection Policies
Var_Sel scheme performs better than Max_Corr selection scheme, i.e., correlation based selection scheme. Maximum correlation based selection algorithm requires scan over all VMs on a host and is repeated for each host. Var_Sel scheme also scans over all the VMs in each host and same is then repeated for each host in data center. The improved performances establish that execution of Var_Sel is faster than slowest selection scheme. Comparison with other schemes was dispensed as most of other selection schemes don’t require to scan all VMs on a given host. Figure 4.9 display the standard deviation characteristics of VM selection execution time. Var_Sel performs better than Maximum correlation selection scheme. All other schemes namely, minimum migration time, random selection and maximum utilization schemes are out of scope while comparing them on this parameter. Standard Deviation characteristics of Var_Sel scheme are better than slowest selection algorithm, i.e., maximum correlation. Low values in comparison to maximum correlation imply that deviation in execution times is less than that of maximum correlation based selection scheme.

Mean time before migration characteristics are displayed in {Figure 4.10}. Var_Sel characteristics exhibit that proposed selection scheme is able to migrate selected VM faster than any other scheme. Our assumption for considering VM’s classification on the basis of bandwidth is rewarding. Mean migration time before migration is almost double in every other selection scheme. The performance of Var_Sel is consistent across all the allocation strategies. This improvement is reflected in SLA Time per Active Host also. SLATAH characteristics of Var_Sel are also better because of this characteristic only. It has contributed further in reducing SLAV characteristics. Figure 4.11 displays standard deviation characteristics corresponding to {Figure 4.10}. This exhibits that Var_Sel has proved consistent performance in VM migration.
Figure 4.12: Host Selection Mean Time Characteristics in VM Selection Policies

Figure 4.13: Host Selection Standard Deviation Characteristics of VM Selection Policies
Figure 4.12 draws host selection mean time. The characteristics are drawn between Maximum correlation (Max_Corr) selection scheme and Var_Sel. Proposed scheme performs better than Max_Corr scheme. Improvement in proposed scheme is attributed to remaining load unification. Figure 4.13 displays that deviation of host selection time with respect to mean host selection time. As host selection time is based on host’s current load, Var_Sel has greatest impact on remaining loading factor of hosts. After migrations hosts are left at almost common load level. This allows quick determination of destination host for migration.

4.7 Summary

This chapter presents a Variance based VM Migration scheme called Var_Sel. Through rigorous simulations, strength of Var_Sel have been presented. An cost effective task scheduling using hybrid approach in cloud implemented in Matlab will be presented in Chapter 5.