Technical Report

UGC Major Research Project on " Building An adaptive Resource Provisioning Framework for applications with Multiple Constraints in Cloud environments "

Principal Investigator - Prof.S.Ramachandram,
Vice-Chancellor,
Professor
Dept. of Computer Science & Engineering,
University College of Engineering,
Osmania University, Hyderabad.

Co- investigator- Prof.P.V.Sudha ,
Professor,
Dept. of Computer Science & Engineering,
University College of Engineering,
Osmania University , Hyderabad

Table of Contents

List of Figures

1. Introduction to the Project :

Cloud computing as a technology and business enabler has been the most accepted change in this decade. The cloud computing definition by NIST (National institute of Science & Technology, Department of Commerce, U.S) highlights core ideas. Cloud computing is defined by NIST as *"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal man- agement effort or service provider interaction "* . Cloud computing is being widely adopted by many organizations because of cost effective solutions offered to users requirements.

An architecture that depicts the various components of Cloud is shown in figure.1.1.

A representation that includes components of Cloud computing Sys- tem shows the four main actors as defined by NIST. These are shown in figure.1.2. Each actor is an entity (a person or an organization) that plays assigned role in cloud.

The cloud consumer is the ultimate stakeholder. A cloud consumer selects the required service, signs feasible contracts and uses the ser- vice. The cloud consumer is billed for the services requested and makes

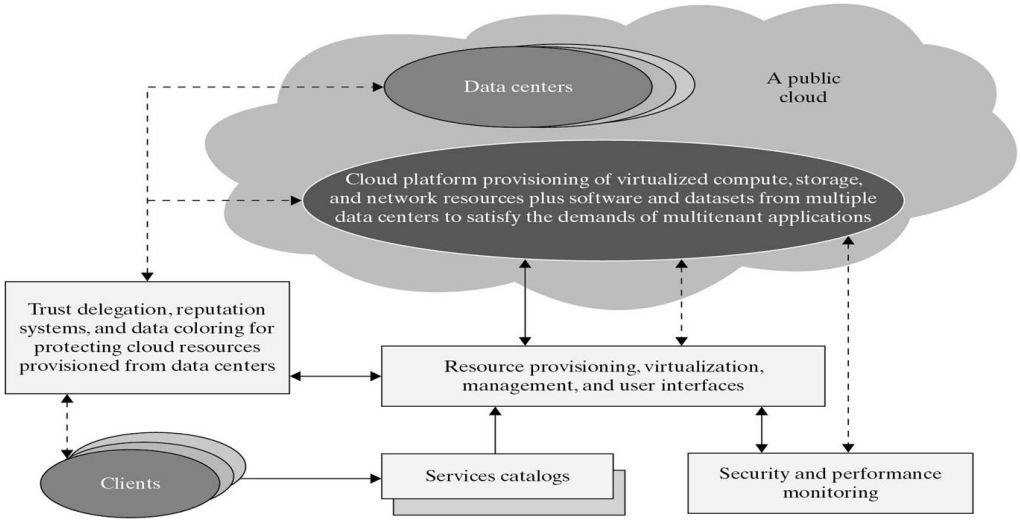


Figure 1.1: Basic architecture of Cloud

payments for the services made available to him. For different types of Service models that Cloud provides like- Software as a service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) some example usage scenarios are shown in figure. 1.3

An excerpt from Gartner Press Release SYDNEY, Australia, May 18, 2015 says *"Global spending on IaaS is expected to reach almost US $16.5 billion in 2015, an increase of*

32.8 percent from 2014, with a compound annual growth rate(CAGR) from 2014 to 2019 forecast at 29.1 percent."

Cloud providers have to play various roles in the process of managing the system. A cloud provider has to provide the requested soft- ware/platform/ infrastructure services, has to manage technical features required for providing the services, has to provision the services at acceptable SLA levels and protect the security and privacy of the services. Cloud service management is a complex task consisting of among various other activities, resource provisioning.

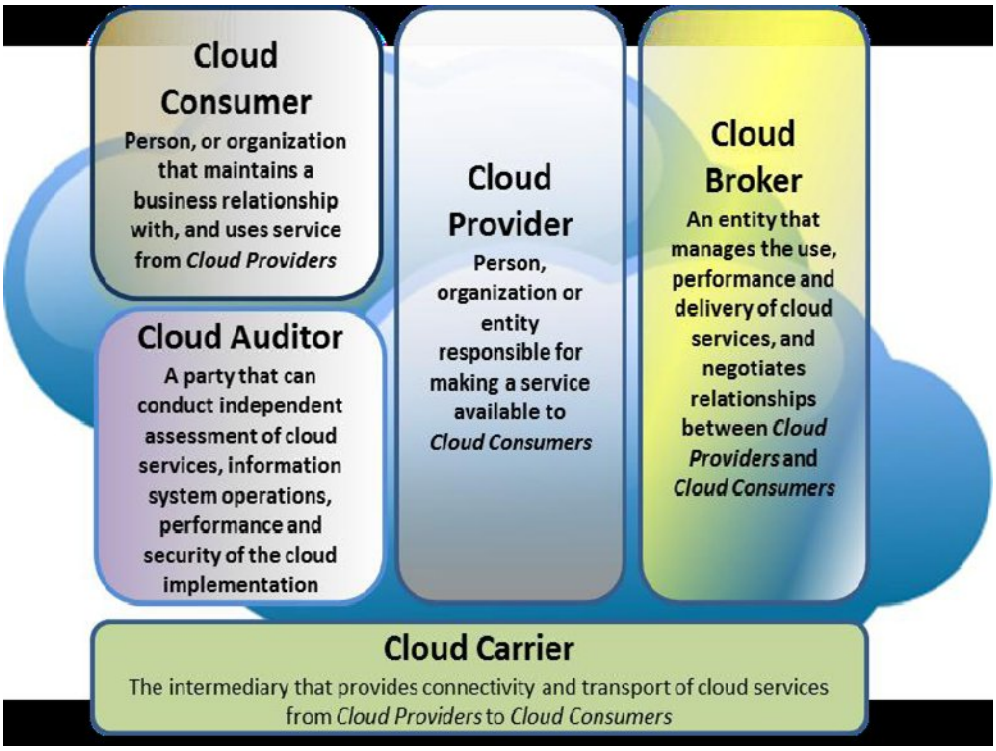


Figure 1.2: Cloud Actors

Currently users are allocated resources based on their requests. Re- sources requested by users are found to be overestimated than their actual requirement. Underestimation of resources can cause resource shortage and consequent revenue loss due to penalties for SLA (Service Level Agreement) violation. Overestimation can lead to idle resources and increased costs. It is necessary to utilize these resources properly in order to decrease cost to each user. Provisioning of re- sources is a challenging issue being faced by the service providers in Cloud, because the requests come from numerous users, requirements dynamically change and there is no specific pattern, trend or seasonality in the resource usage of users on Cloud. The Google cloud usage data published as trace has been studied and the usage of resources and resource requested in shown in figure 1.4

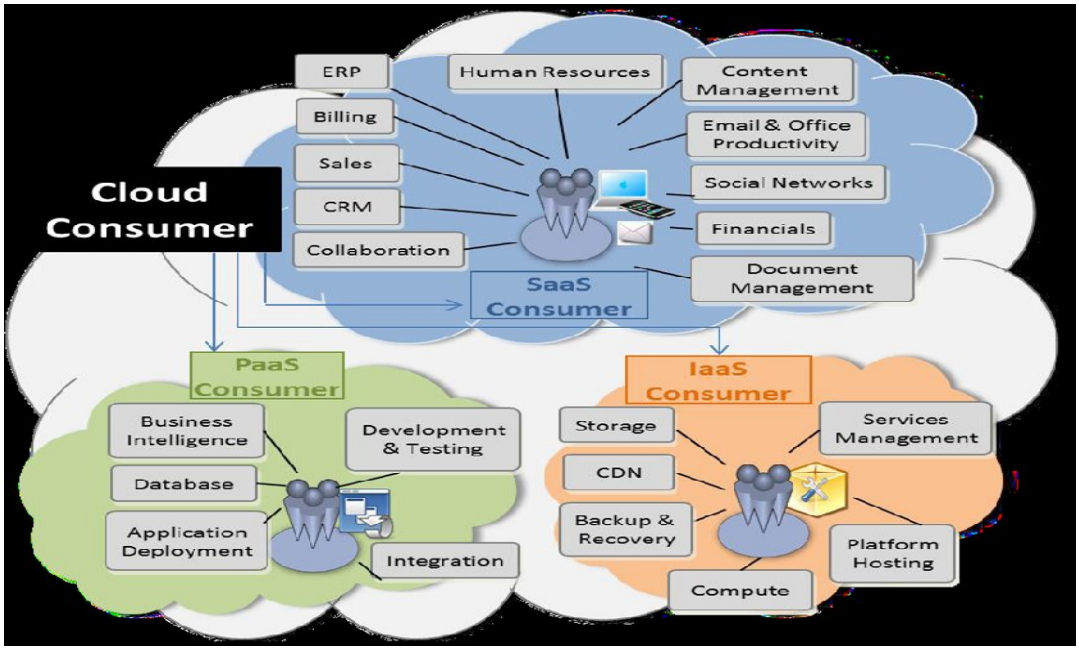


Figure 1.3: Cloud Consumer use case scenarios

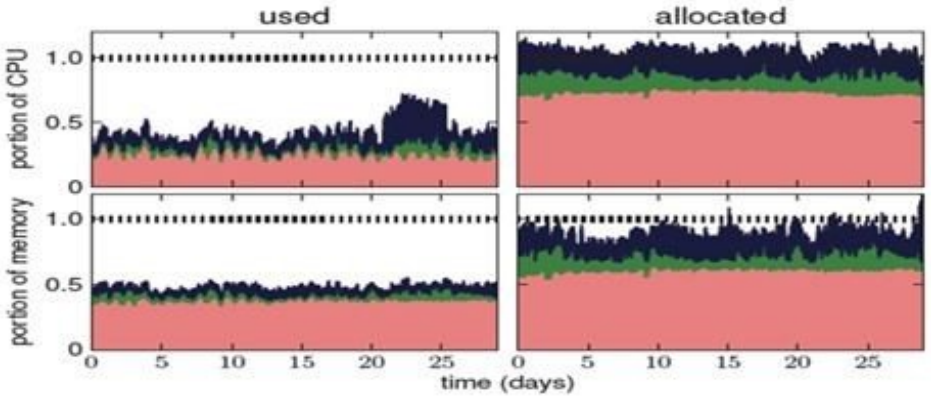


Figure 1.4: Used Resource Vs allocated resource of Google Cluster Data.

This shows that, resource requirements of the users need to be met as per their use rather than what users request for. The real challenge is to be able to provide to the users such quantity of resources that they will actually use.

Data Centers are used to house the large compute and storage resources to be made available to users by the cloud providers. Cloud computing applications run over multiple computers connected by a network .

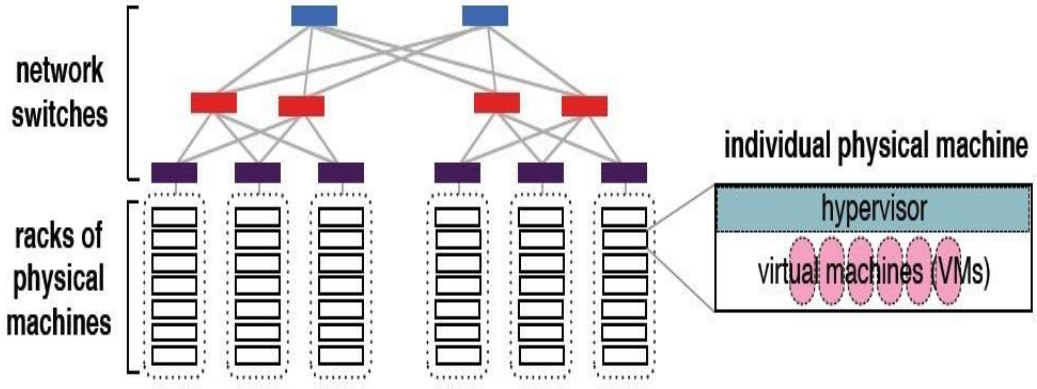


Figure 1.5: Typical data center network

As can be seen in the figure 1.5 each of the various virtual machines that are available, resources will be provisioned to them. How resource management relates to power consumption and power saving is very important for modern day businesses. Google is the first major web company that has revealed its power consumption information. Google uses 260 million watts continuously across the globe as mentioned in their blog-Google Green. According to Koomey, data centers comprised 1.3% of the global energy usage in 2010. Similarly other major cloud providers use huge amount of power. Assuming three-year server and 15-year other infrastructure cost consolidation, energy-related costs are estimated to amount to 41.6 % of operation cost of large-scale Data Centers as discussed by Hamilton in and as shown in figure. 1.6 .Only power costs individually, are lower than infrastructure costs, and less than the servers themselves. Servers are the dominant cost. Power is only 23% of the total, but power distribution and cooling make up 82% of the costs of infrastructure. Cost of building is 12-15%. Hence, overall power consumption costs are considerable.

2. Objectives of the Project :

Most of the organizations currently heavily rely on using Infrastructure, Platform or Software as a Service from various Cloud Providers. These Providers expect users to give request for resources that they require and are billed accordingly. The users of cloud resources do not want to have business discontinuities due to unavailability of resources throughout their business process. This results in over-estimated request for resources. An adaptive system the can predict the resource requirement of users and also automatically scales as opposed to reactive scaling will enable making huge resource

saving for the Cloud Resource Providers and saving for the users as they will be billed only for their used resource and the resource wastage will drastically reduce. In all this process, meeting users' Service Level Agreements is important pre-requisite. The main objectives of this project are:

1. Developing a framework which manages to provide resources within stated SLAs
2. Developing Adaptive system that manages resources and services with reduced preparation / setup time .
3. Developing a system by having SLAs specified explicitly and ensuring that the system meets them

2. 1. Whether Objectives were achieved :

Yes, The stated objectives were achieved.

An adaptive Framework that provides Resources within stated SLAs, enables Predictive Auto-scaling was developed.
Additionally, Energy Savings achieved because of the Prediction were also computed for two implementations - Intel Xeon E5 2687 and Intel Xeon E5 2697 , using approaches Distance Weighted Averaging (DWA) and Locally Weighted Regression (LWR) .

3. Challenges :

Resource provisioning and Energy estimation gets complicated in the Cloud scenario as compared to the Web based and Grid based systems because:

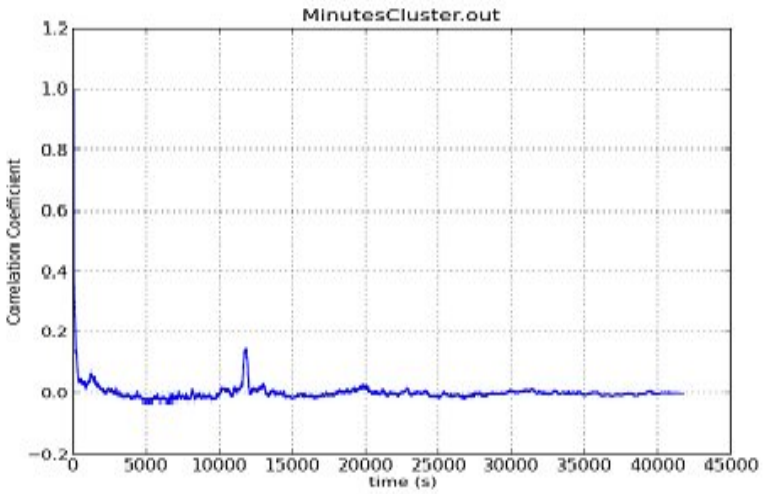
a) Resources are requested by the user in real time. The resources are requested when the application starts and this information is not available beforehand. Users **expect Instantaneous Resource availability**. The resources are to be made available to the users as the application execution proceeds. This is difficult to implement by the Cloud Provider because of erratic requests of various users as shown by the Cloud usage trace - Google Trace data as described in [13].

b) To make available resources as per changing requests of users, to enable **dynamic scalability of resources**, reactive approaches are easier to implement but take unacceptable time to provision resources. Predictive approaches serve the purpose of resource reservation but are difficult to implement as they depend on historical data and on cloud. New users without historical information are more common, making proactive provisioning difficult for cloud environment.

c) The trace of Cloud usage show that it does not lend itself to existing prediction approaches like Time series, Queuing models Bayesian model, SVM, Neural networks etc., because of absence of patterns, trends and seasonality. The ACF(Auto Correlation Function) study of Google Cluster Data from figure. 3.1 and as shown in figure. 3.4. **Cloud**

access and hence resource usage shows no specific pattern (*From [14]*). Corresponding correlogram is shown here in fig 3.1.

Figure 3.1: Correlogram of Google ClusterData



d) The machines that are made available to users by various providers like Google Cloud Platform- Google Compute Engine - available at cloud.google.com/compute/docs are quite variable. For e.g. Standard machine types, High memory machine types, High CPU machine types as shown in fig. 3.2

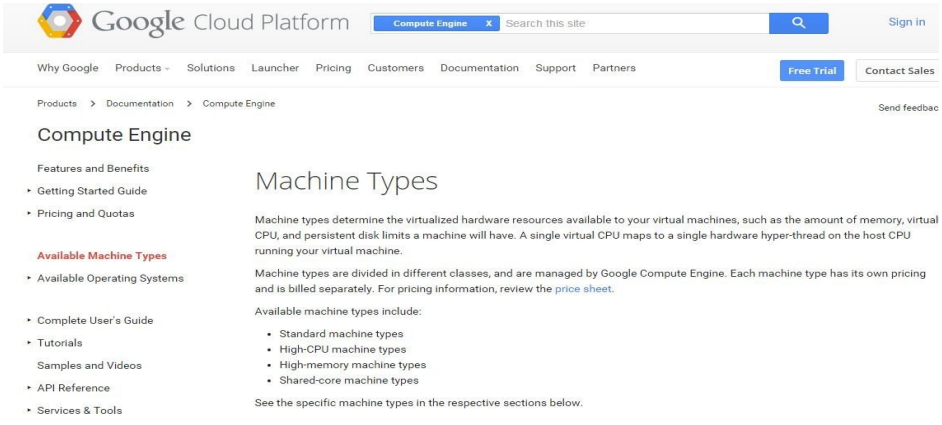


Figure 3.2 : Various machine types on GCE

Detailed information for high CPU type machines available from the website shows how various types of resource requests need to be met for different users tasks. This snapshot in fig. 3.3 shows the details.

e) **Requests for these machines compute units, the virtual CPUs are actually implemented on Intel Sandy Bridge, Intel IVY Bridge, Intel Haswell machines**. Google Compute Engine unit is a unit of CPU capacity that is used to describe the compute capability of machine types. Google has chosen 2.75 GCEUs to represent the minimum computational capacity of one virtual CPU (a hardware hyper-thread) on Sandy Bridge, Ivy Bridge, or Haswell platforms. **These are different in their energy consumption--idle, full load and average. Therefore it becomes difficult to exactly estimate the power consumed per user request for CPU**. Recent work by Ian Cutress at Anandtech helps in knowing these power consumption as shown in figure.3.4 .

f) . Energy consumed per job, per task cannot be estimated as an absolute number because of the data obfuscation in Google Cluster data. Dynamic power consumption here based on the resource units consumed gives us energy consumed and energy saved on two models - Intel Xeon 2687 and Intel Xeon 2697 is a relative measure. Same approach can be employed on various other machines where proportional increase in power consumption by use of compute units is known.

High-CPU machine types

High-CPU machine types are ideal for tasks that require more virtual CPUs relative to memory. High-CPU machine types have one virtual CPU for every 0.90 GB of RAM.

Machine name	Description	Virtual CPUs[1]	Memory (GB)	GCEUs[2]	Max number of persistent disks (PDs)[3]	Max total PD size (TB)
n1-highcpu-2	High-CPU machine type with 2 virtual CPUs and 1.80 GB of memory.	2	1.80	5.50	16	10
n1-highcpu-4	High-CPU machine type with 4 virtual CPUs and 3.60 GB of memory.	4	3.60	11		
n1-highcpu-8	High-CPU machine type with 8 virtual CPUs and 7.20 GB of memory.	8	7.20	22		
n1-highcpu-16	High-CPU machine type with 16 virtual CPUs and 14.4 GB of memory.	16	14.4	44		
n1-highcpu-32[4] (Beta)	High-CPU machine type with 32 virtual CPUs and 28.8 GB of memory.	32	28.8	88		

Figure 3.3 : Machine types for Google Compute Engine

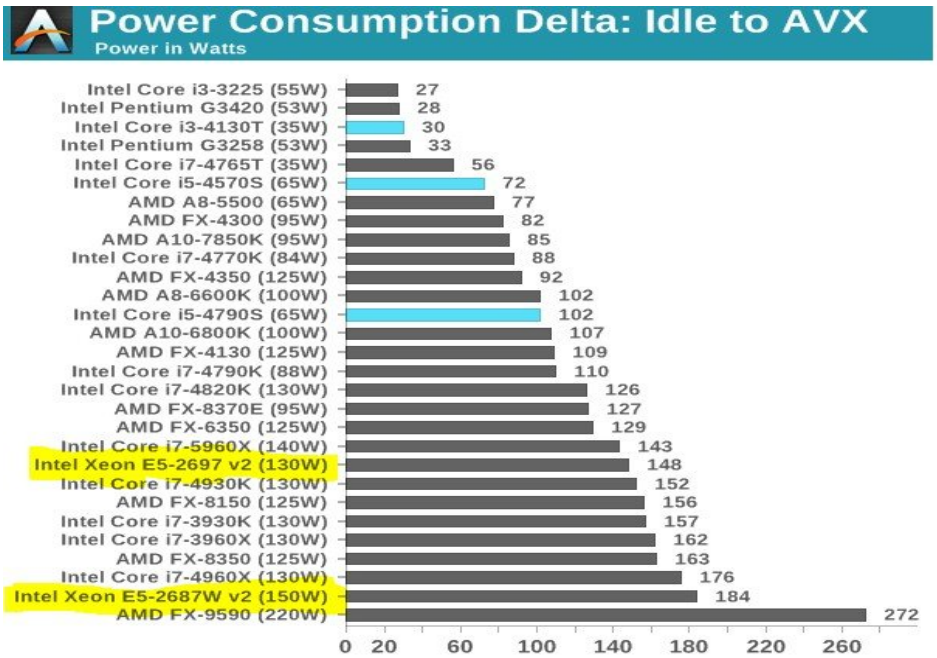


Figure 3.4: Power consumption for two chips of IVYBridge

Work related to Real Cloud usage data is studied to find if any work related to prediction approaches using real cloud data, at least simu- lated data of real cloud usage is done. Real cloud usage data has been released by Google in 2011. The trace ver 1.0 is only a short trace that describes a 7 hour period from one cell (cluster). We have used the Google Trace Ver 2.0 , which is a 29 day trace on 12k-machine cell in May 2011 with following statistics:

Trace Span	29 days	No. of Servers	12,538
No. of Tasks	17,75,951	Avg. Tasks/day	6,12,170
No. of Users	430	Avg. Users/day	153.20
		Avg. Tasks/user	3,981.06

Figure 3.5: Consolidated information regarding trace

The work using this trace is useful to underline the scale of resources, users and hence the requests that need to be focused in developing the prediction algorithm for real cloud usage data. The paper by Mishra et al., [81]captures the heterogeneity and dynamicity of data under consideration the Google trace. In [1] by Charles Reiss, Consolidated cloud environments are constructed from numerous machine classes. Extremely dynamic resource demands with high variation over short duration are observed. This leads to various issues like rapid task scheduling decision making, revision of previous assignment decisions, prediction interference for resources over time. Heterogeneity is observed in Machine types and attributes, workload types, job duration, task shapes and distributions. Dynamicity is observed in Machine churn, task

and job churn.

In [82], authors compare the two workloads- GridMix3 and Yahoo production cluster by using k means clustering approach.

By using time series the data of real cloud usage trace from IBM hosted cloud is studied for both CPU and Memory usage is [83]. By using the Principal Component Analysis approach, prediction for resource usage across multiple nodes is implemented. Using these predictions and spatial correlations across cluster of virtual machines, better utilization of physical machines is suggested.

Synthetic workload was generated in [84] by using the resource utilization and task wait time. Deriving characterization models of task usage shapes from Google compute cloud of 6 clusters spanning 5 days is used. This is smaller trace as compared to what we are using. An important paper that discusses about scheduling and dependencies is [85] by Sharma. In order to identify the performance of compute clusters at Google scale, realistic workloads are needed. So that Task placement constraints as used in actual Google Clusters are to be incorporated into workloads apart from the resource requirements. A methodology Consisting of data preparation workflows, baseline work- flows and treatment workflows is used. Machine constraint characterization and task constraint characterization are obtained. This is used in identifying performance impact of task placement constraints.

A useful paper that helps in understanding the workload (for jobs, tasks) and host load (at machine level) in a Google Data Center in comparison to the Grid system is provided in [86] by Sheng Di. Comparison of Workload of job or task length of job or task, job priority, job submission frequency and job resource utilization are discussed. For host load, maximum load and machine usage level between grids and clouds have been shown. This kind of characterization is very useful for job scheduling and prediction of load on various machines . In [87] authors have provided a reusable approach for characterizing cloud workload obtained from the Google Compute trace of 29 days. The characterization has considered patterns for both the users and tasks. Their characterization across various clusters helps researchers actually use the published trace more specific to

their application.

By using an approach called Autonomous cooperative cloud based platforms (ACCP), in [88], authors have shown the advantage of co- operative approach in Inter-Cloud environments. A simulation based evaluation based on SCIENCE CLOUD under a model of realistic Google Cluster data set [89] has been used. They have tried to show that this approach gives better performance as compared to selfish approach in large clusters.

3.1 Issues with existing prediction techniques

Actually, study of related work started with trying find if efficient scaling techniques are available. In current practice, cloud scaling is still reactive. Reactive rule based methods enable scaling based on a specific metric reaching predefined limits. Reactive techniques add re- sources to the system on happening of certain events like performance degradation, certain time being elapsed or thresholds on number of users, number of active VMs available at certain point in time. This approach is implemented by several cloud providers such as Amazon [7], in third party tools such as RightScale [11] or AzureWatch [35] Predictive scaling shown an encouraging alternative to reactive scaling specifically when we expect resources to be available beforehand and do not want the users to wait. The other advantages and limitations of predictive scaling are discussed in next section. Existing body of research work focuses on various types of predictive techniques to tell how much of resource will be required by the various users in next period of time. These techniques are discussed in detail in Section no. 2. A study of the existing approaches to prediction, available data traces that have been used in prediction has been done (as listed in subsection 2.4). To arrive at a solution, most of the available approaches have been weighed for feasibility. Based on current research body of work available following conclusions are reached :

1. Most of the techniques used today in resource prediction are based on prediction for Web workloads. **The main difference between Web workloads and Cloud Workloads is the absence of trend, seasonality and**

diurnality in Cloud workloads. Hence, the approaches used in Cloud resource prediction need to be different from those for Web workloads.

2. Current approaches to predict resource requirements are based on availability and processing of large amount of history data. This is not necessarily possible when the prediction is to be done for initial start cases in any system.
3. Fetching and Processing of history data requires time even with best algorithm and data access techniques. **This time lag is dif- ficult to afford for real time prediction in Cloud resource pro- visioning**
4. Some of the combined approaches shown by other researches need to select specific approaches based on available data. But this would also show some lag time and not enable prediction in real time.
5. The work on energy saving has mostly been coarse grained, whereas our approach is more fine grained. We inspect the possible energy saving that can be obtained by saving resource units, more specifically CPU units for each job of individual user.

3.1.1 Cloud mining: Resource Requests and Usage

The problem of low utilization of resources can be overcome by good prediction techniques. Bridging the gap between real requirements and estimated requirements of the users, will lead to huge savings for both the providers and also to the users. Indirectly the sum total of saving that could be derived per user taking his various jobs, multiple tasks and multiple days on which he makes these requests, will be huge. This huge amount of resource saving is also accompanied by huge amount of energy savings. Per user savings are calculated here. This can be extrapolated to see the data center wide savings that can be achieved per month. This in real time is a significant contribution of our work.

The scatter plot in fig. 3.6 is representation of CPU requests and CPU usage of tasks. There is a large variation in users resource requests.

Observations from the study of the Google cluster data can be summarized as:

1. Performing analysis on large-scale trace-logs is fundamental to deriving realistic prediction models.
2. The Cloud environment does not exhibit obvious cyclic behavior. Patterns of cloud load data are not cyclic, seasonal, diurnal. Cur- rent load is not related to previous load observations. This is the main reason for failure of most of the existing approaches based on Time series, Hidden Markov model, Bayes method etc.

3. Workloads in clouds are highly variable with respect to time.

4.Users grossly overestimate the resources required to meet business objectives. This forms the premise for this work.

5. Modeling user behavior is a critical factor when characterizing Cloud workloads. User behavior is important to predict future resource requirements.

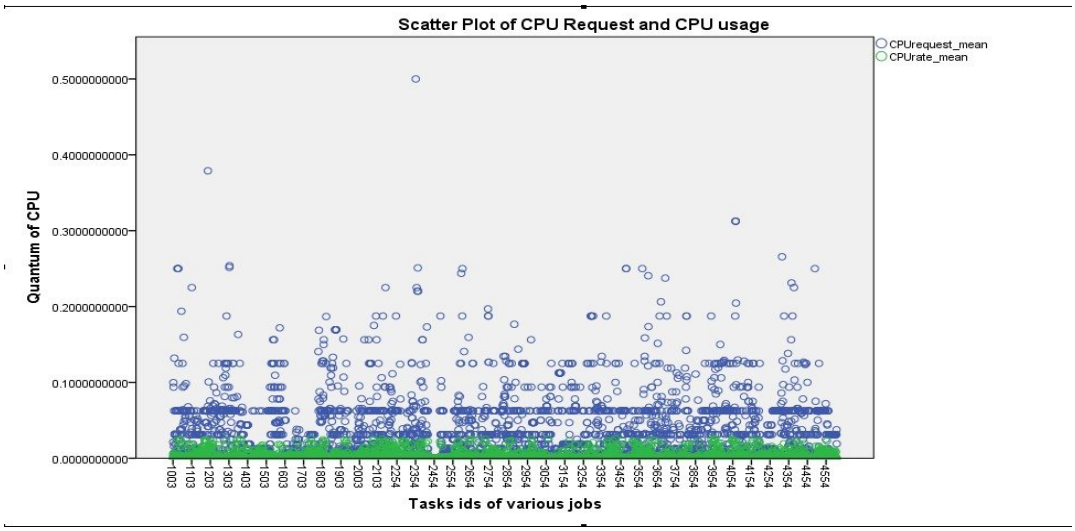


Figure 3.6: Scatter plot of the CPU requests and CPU usage values for bucket 0,bucket 1 of Google cluster data

3.2 Enhanced Instance Based Learning

In implementation of Enhanced instance based learning, from the stated trace data , we have used user id, CPU requested, CPU used, memory requested, memory used for various cases. Resource usage prediction for the user for whom only his amount of **resource requested** information is available is computed. What would be the energy saved because of each variant of resource prediction is also made available.

Such estimated values for CPU Usage and Memory Usage are predicted Compute units and Memory units that will be required by Cloud user in the near future. Estimated energy saved is also estimated for two different machine types - (INTEL XEON E5 2687 and INTEL XEON E5 2697)

4. Achievement of the Project

Significant achievements resulted from this work. The Resource Provisioning Framework that is necessary had to take into consideration multiple constraints like Users' Service Level Agreements which are interdependent.

1. In trying to have a scalable system, it became necessary to get a resource prediction system that would overcome the limitations of the existing approaches to scaling like time delay because of reactive scaling , wasted resources for the providers and unnecessary cost for the Cloud users.
2. Proposed work uses Enhanced Instance Based Learning approach for resource requirements prediction and runs this approach with actual Cloud trace data. The energy saving that would be achieved with predicted resources as compared to existing user requests for resources are shown.
3. The entire solution is tested by having a real private cloud using HP Cloud system Enterprise installed on our servers in the Cloud computing Lab.Using the proposed Enhanced Instance Based Learning (EIBL), the saving of resource units per collection of machines per hour is quite significant. This also translates to significant saving of energy units per user jobs. This enables better and efficient utilization of the resources by the service provider.
4. This proposed approach when scaled and worked in real cloud system, are shown to be extremely beneficial to both - the users and service providers.
   Results obtained from proposed predictive approach shows Resource saving (compute and Memory), energy saved, accuracy of prediction and data size used by prediction. The stated SLAs of the Users are adhered to. The required Auto-scaling is shown to be efficient.

There is a considerable improvement over existing ad-hoc approach by the users. Specific contributions can be enlisted as:

a) An enhanced approach for prediction of resources requirements , specifically in Cloud environment is developed. When DWA(Distance Weighted Averaging) and LWR(Locally weighted Regression) are employed for prediction, the savings of resource units per job of a user are shown in results. Resource units saved per job as percentage of resource saved compared to what was actually getting wasted by ad-hoc method of request for resource is found to be on average - 90.68% saving for CPU estimates using DWA, 73.44% saving for CPU estimates using LWR, and 73.28% saving for Memory estimates using DWA, 88.47% saving for Memory estimates using LWR.

b) The percentage reduction in resource estimation error is on average- 95 % for CPU estimates using DWA,92.5 % for CPU estimates using LWR and 67.13 % for Memory estimates using DWA, 47 % for Memory estimates using LWR.

c) The resource units saved by DWA and LWR Compute units pre- diction approach also helps in computing Energy saved on INTEL Xeon E5 2687 and INTEL Xeon E5 2697. Energy saving computation for various jobs shows average saving of 63% .

d) The proposed approaches for prediction use small amount of data for processing to arrive at the result. Use of *K means* achieves this objective. Hence, it requires less time to process the prediction requirements. This is an important consideration for Cloud systems where dynamic re- configuration based on users requests and actual use of resources will play a crucial role in successfully incorporating this prediction approach into real cloud provider system in the future.

e) The framework built takes as input the users SLAs, then scales to the requested resources as per SLAs , internally using the EIBL Prediction approach for the implementation. This makes it possible for us to show that SLAs can be met by Cloud Providers more easily as compared to present Cloud Providers who have to waste resources being overbooked and not utilized by hundreds of users at the same time.

Finally, the results of prediction of resources have been executed and tested with real time data trace of Cloud.

A Private cloud is set-up in the college to enable testing of this framework . HP servers - **HP DL 380 G9 Server + HP 2920-24G-POE Switch – 3nos. ,**with Virtual connect and HP matrix operating environment have been procured Software component - HP Cloud system Enterprise with Cloud system Automation was **procured and installed**

This platform gives cloud service providers and enterprises the ability to manage the complete lifecycle of their cloud service products. The platform enables product creation based on service templates, which are generated by utilizing the underlying HP Cloud System software and hardware.

HP Cloud System Enterprise enables the distribution, subscription, and consumption of on-demand cloud services and other IaaS and hosted services, as well as third-party SaaS services.

Service Level Agreements are important part of the framework that ensures that Users of the Cloud are satisfied . Hence, this is an efficient approach to Capacity Planning for large data centers.

5. Summary of the Findings :

The Cloud usage scenario has various service providers and large number of users. This makes it useful to provide Framework for resource provisioning when there are multiple constraints like - Task completion time(deadlines) , cost , easy scaling etc.

1. The dynamic demand for resources from Cloud makes resource management extremely important in design and decision-making processes in cloud computing environments. Providers of resources on Cloud offer heterogeneous resources such as compute units, memory and storage in Virtual Machine instances (VM). Large scale data centers are essential to service the

huge rise in demand for reliable high performance computational and storage services over cloud.

2. From real traces, resource requests of the users on cloud show that, they are mostly overestimated and sometimes underestimated . Over multiple requests in each task, multiple jobs, this cumulative unutilized resource amounts to considerable wasted resources for provider and unnecessary expenditure for users. Possible approaches and techniques that can ensure that resources are not overbooked and wasted are an urgent necessity.

With increase in computational and storage needs of business users who are moving their applications and data to cloud, scaling of cloud resources at the providers end, results in increased energy consumption and carbon dioxide emissions. This necessitates looking for possible energy saving approaches be adopted in data centers

.

3. Mega data centers that house thousands of servers and consume huge energy per hour at peak times lead to increased operational costs. Power consumption contributes to 42 % of data centers monthly expenses. More important consideration is that, the huge power consumption hastens climate change due to carbon dioxide emissions and use of nonrenewable energy sources. Therefore, for environmental and financial reasons it is imperative to try to reduce unnecessary booking of resources when they are not actually used.

4. This problem can be overcome by finding suitable resource requirement prediction approach, such that, based on predicted required amount of resource, resource reservation will be done. Efficient resource requirement prediction approach will ensure that resources are efficiently utilized. Resource usage studied from existing workloads on cloud, more specifically the Google Compute Cloud usage data, shows that these do not have any specific pattern, trend, seasonality in the use of resources. Various researchers who have used resource prediction techniques such as time Series, exponential smoothing, neural networks, Bayes method etc., have shown results for only Web based distributed data, which is much different from actual cloud usage data. Cloud Usage data shows bursty non cyclic and non-seasonal behavior. This makes resource requirement prediction for Cloud Challenging.

5. The proposed prediction approach applies Enhanced Instance Based Learning(EIBL approach. Two variations of EIBL - Distance weighted averaging(DWA) and Locally weighted Regression(LWR) have been used in experiments with Google Cluster Data. Google Cluster Data is a trace of the data used by various users of Google Cloud over a period of 29 day. The proposed prediction approaches are tested using Google cluster trace data.

6. The results show considerable saving of resources and energy. Resource savings obtained are - average CPU saved is 79%, average Memory saved is 60% and average Energy saved is 60%. Resource predictions obtained by this work are very close to what is actually used by the user. The proposed approach has exhibited high accuracy of 99% as compared to resource usage values from Google trace data. This underlines the contribution of this work to existing body of work on cloud resource management, energy saving approaches and green computing as a

consequence. The framework that is developed around this, uses the Service Level Agreement requirements of the users and scales upto the requirements in real time.

6. Contributions to the Society :

Cloud adoption is a major change in the recent times . There are important issues and solutions are required for them.

1. Cloud users are not aware of the resources that actually get wasted, when they request for resources by only making an estimate of their needs. There are a number of businesses which try to use the compute and storage from cloud to meet their end users needs. For eg. online shopping sites like Amazon , Flipkart etc., store their product information on Cloud and enable retrieval of specific Product related information when customers make request. They try to overbook resources expecting scaling of business. **This leads to huge loss of booked resources and hence they end up paying much more than they actually use. With this developed solution, we are able to show huge saving for each instance of his resource usage. The cumulative saving over multiple such requests, over multiple months, it would show huge savings for each customer.**

2. For the Cloud Providers, the resources that were overbooked but not used are wasted because they are blocked. With the proposed solution the resources need not be provided to uses based on their estimate. They can be provided by using the Prediction technique proposed. This results in allocation of only resources that each user would use**. So, huge number of resources are saved which can actually be provisioned to the other customers. So, the number of unique requests that can be simultaneously services by the Cloud Provider is increased.**

3. The time required for bundling the VMs as per requests of users, then provisioning them to customers results in considerable wasted time as in case of reactive scaling used by Amazon auto-scaling and Salesforce auto-scaling. Cumulative wastage of time would be high. **So, a predictive scaling technique that enables pre-bundling of resources that user would require in the future , ensures that there is no need for the users to make requests, and bundling time would also be saved.**

4. Capacity Planning is a major challenge in large organizations because of non availability of specific trend, diurnality and pattern. Though other approaches for prediction like time series using AR,ARMA, ARIMA, Bayes method, Neural Networks etc., fail for Cloud usage trends, this proposed prediction has been proved to be much nearer to actual cloud resource usage. This ensures that better capacity planning can be done while also living upto SLAs of the customers.

5. The rapid use of Internet-enabled devices, use of Software-as-a-Service, growth of Big Data applications and organizations moving their operations to Cloud-Based Systems has accelerated the growth of large scale Data Centers. The Energy used in a large scale data center is a matter of great concern for all organizations and nations which are concerned about climate change. Some of the largest and most complex data centers are owned by well-known internet giants like Facebook, Google, and Amazon etc. Organizations strive to reduce operations and energy costs by implementing a variety of solutions such as reliance on renewable energy sources and power saving in data centers. For a normal user of cloud, services available from providers like Google Cloud Platform, Amazon Web services etc., upfront capital costs have dramatically reduced and has enabled them to scale quickly. To be able to provide such services to the users in real time, the operational challenges in managing a Data Center are huge. One of the most complex challenges is Energy conservation. Energy efficiency is a very important consideration world over. Large organizations have focused on identifying ways for fostering energy-efficient protocols, architectures and techniques. To ensure maximum availability of services to the clients, most of the data centers try to provision more resources so that customers get the resources 24/7 and do not have a chance to complain. Most challenging research is happening on energy consumption levels of data centers. The Environmental Protection Agency (EPA) study showed that energy usage doubled from 2006 (61 billion kWh) to 2011. In 2009, data centers used 2% of worldwide power usage at expenditure of US $30 billion. Gartner had forecast cloud related expenditure for 2012 to be at $106.4 billion, a 12.7% increase from 2011, and revenue is forecast to change from $163 billion in 2011 to $240 billion in 2016 . The IT sector contributed to 2 % of carbon dioxide worldwide in 2005, and it is growing by 6% per year. Nearly 80% of energy costs can be reduced by effective measures on various devices in a data center and 47 M metric tons of carbon dioxide emissions per year can be reduced.

The selected trace, Google Cluster data, is a trace of users resource requests for Google compute cluster. Google Compute Engine unit is a unit of CPU capacity that is used to describe the compute capability of machine types. Google has chosen 2.75 GCEUs to represent the minimum computational capacity of one virtual CPU (a hardware hyper-thread) on Sandy Bridge, Ivy Bridge, or Haswell platforms. Hence, as a example case of implementation on IVY Bridge for the two processors, Intel Xeon E5 2687 and Intel Xeon E5 2697 Energy saved is calculated based on idle Thermal Dissipation and Maximum energy consumed. Energy consumed per job, by the user in each method - adhoc resource request method, by using DWA, by using LWR, for each type of machine Intel Xeon E5 2687 and Intel Xeon E5 2697 **is computed.** The Resource requests in trace are a fraction of maximum resource in the trace. $Energy\ range = (Max\ energy(on\ full\ loading) - Min\ energy\ (on\ no\ load))$

$$For\ Intel\ XeonE5\ 2687, Energy\ range = 184W - 150W = 34W$$

$$For\ Intel\ Xeon\ E52697, Energy\ range = 148W - 130W = 18W$$

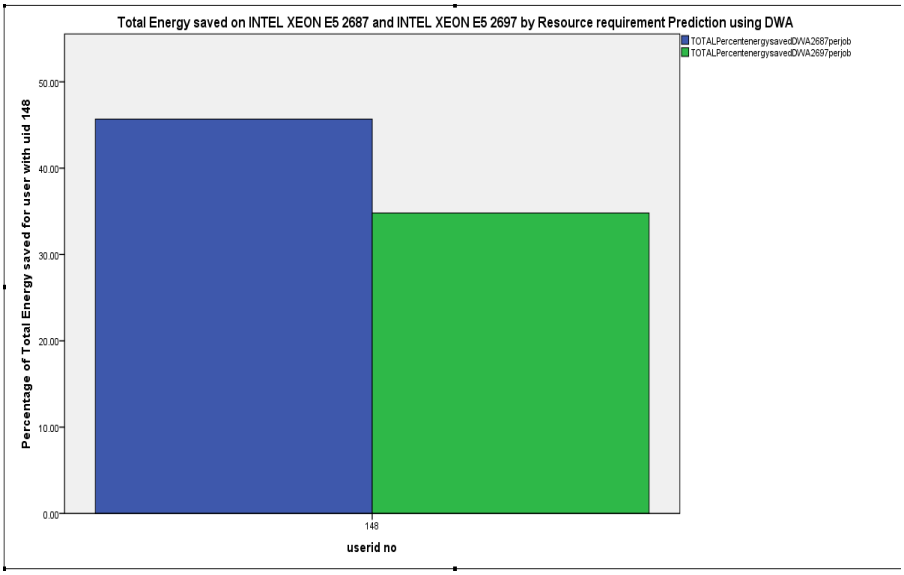


Fig. 6.1: Percentage of Total Energy saved **per user** (sum of energy savings of all jobs of each user),On Intel XEON E5 2687 and Intel Xeon E5 2697 due to resource requirement prediction by DWA for the user with uid 148

Similarly , Energy Saved per user for execution on Intel Xeon E5 2687 and Intel Xeon E5 2697 using LWR method are also computed.

Both these methods- DWA and LWR have shown energy savings of nearly 50%, which amounts to huge value when cumulative savings of energy are taken into consideration.

This is a major contribution to the society.

7. Comprehensive Results obtained :

The tables below give a comprehensive information regarding the savings obtained, accuracy of prediction and the energy savings obtained by this work using two approaches- DWA(Distance Weighted Averaging) and LWR(Locally Weighted Regression) .

	Observed variation as compared with Actual CPU Usage		Percentage reduction in variation	Saving of Compute units
	For Initial Adhoc estimate for resources. (CPU request – CPU Usage) (A1)	For Prediction by DWA (CPU predicted by DWA – CPU usage) (A2)	((A1 – A2) / A1) * 100	Resource units initially requested – Predicted resource by DW
Uid = 148	3.8562	0.24979	93.5224	4.1060
Uid =169	14.980	2.7253	81.8071	12.2526
Uid= 412	370.3162	12.1423	96.7212	382.4585

Fig. 7.1. Sum of compute units saved per job by using DWA

	Observed variation as compared with Actual Memory Usage		Percentage reduction in variation	Saving of Memory units
	For Initial Adhoc estimate for resources. (Memory request – Memory Usage) (A1)	For Prediction by LWR (Memory predicted by LWR – Memory usage) (A2)	((A1 – A2) / A1) * 100	Resource units initially requested – Predicted resource by LWR
Uid = 148	0.6280404	0.0635051	89.8876	0.5645353
Uid =169	13.4812	3.1633	76.5355	10.317825
Uid= 412	43.650755	0.480657	98.8988	43.170098

Fig. 7. 2. Sum of compute units saved per job by using LWR

Caseid	For CPU		
	MAPE in User CPU Estimate	MAPE in prediction by DWA	% reduction in error
For Uid= 148	2357.85	291.81	87.6239
For Uid= 169	5341.19	4.694	99.91
For Uid= 412	2976.85	63.615	97.83

Fig. 7.3. Mean Absolute Percentage Error and reduction for CPU estimate per Job by using DWA

Caseid	For CPU		
	MAPE in User CPU Estimate	MAPE in Prediction by LWR	% reduction in error
For Uid= 148	2357.85	255.35	89.17
For Uid= 169	5341.19	307.21	94.25
For Uid= 412	2976.85	173.44	94.17

Fig. 7.4. Mean Absolute Percentage Error and reduction in error for CPU estimate per job by using LWR

Case id	For Memory		
	Error in User Memory estimate	Error in prediction by DWA	% reduction in error
For Uid= 148	69.5218	47.77	37.71
For Uid= 169	173.6548	57.3372	100.71
For Uid= 412	209.9911	77.7577	62.97

Fig. 7. 5. Mean Absolute Percentage error and reduction in error for memory estimates per job by using DWA.

Case id	For Memory		
	Error in User Memory estimate	Error in prediction by LWR	% reduction in error
For Uid= 148	69.5218	56.2085	21.18
For Uid= 169	173.6548	104.8720	49.39
For Uid= 412	209.9911	61.72	70.61

Fig.7.6. Mean Absolute Percentage Error and reduction in error for Memory estimate per job by using LWR

FOR DWA on Intel Xeon E5 2687		Observed Variation as compared to Energy actually Used		Percentage reduction in variation	Saving of energy units	Percentage saving of energy
	Energy consumed with Adhoc requests (A3)	Energy used with Adhoc requests – Energy used with actual resource used (A1)	Energy used with DWA prediction – Energy used actually by user (A2)	((A1 – A2) /A1) * 100	Energy used with Adhoc Requests – Energy used with DWA prediction (A4)	(A4/A3) *100
Uid= 148	305.57	131.11	-8.49	93.5168	139.61	45.69
Uid =169	1069.06	509.25	-17.06	96.6499	526.31	49.23
Uid = 412	13592.35	12590.75	-412.84	96.7210	13003.59	95.67

Fig. 7.7. Energy saved per job by using DWA in 2687

FOR LWR on Intel Xeon E5 2687		Observed Variation as compared to Energy actually Used		Percentage reduction in variation	Saving of energy units	Percentage saving of energy
	Energy consumed with Adhoc requests (A3)	Energy used with Adhoc requests – Energy used with actual resource used (A1)	Energy used with LWR prediction – Energy used actually by user (A2)	((A1 – A2) /A1) * 100	Energy used with Adhoc Requests – Energy used with LWR prediction (A4)	(A3 /A4) *100
Uid= 148	305.57	131.11	-13.97	68.1449	145.08	47.48
Uid =169	1069.06	509.25	-334.63	34.2896	843.88	78.94
Uid = 412	13592.35	12590.75	-419.66	96.6669	13010.41	95.72

Fig. 7.8. Energy saved per job by using LWR in 2687

FOR DWA on Intel Xeon E5 2697		Observed Variation as compared to Energy actually Used		Percentage reduction in variation	Saving of energy units	Percentage saving of energy
	Energy consumed with Adhoc requests (A3)	Energy used with Adhoc requests – Energy used with actual resource used (A1)	Energy used with DWA prediction – Energy used actually by user (A2)	((A1 – A2) /A1) * 100	Energy used with Adhoc Requests – Energy used with DWA prediction	(A4/A3) *100
Uid= 148	212.36	69.41	-4.50	93.5167	73.91	34.80
Uid =169	616.56	269.60	-9.03	96.6505	278.64	45.19
Uid = 412	7246.48	6665.69	-218.56	96.7211	6884.25	95.00

Figure 7.9.: Energy saved per job by using DWA in 2697

FOR DWA on Intel Xeon E5 2697		Observed Variation as compared to Energy actually Used		Percentage reduction in variation	Saving of energy units	Percentage saving of energy
	Energy consumed with Adhoc requests (A3)	Energy used with Adhoc requests – Energy used with actual resource used (A1)	Energy used with DWA prediction – Energy used actually by user (A2)	((A1 – A2) /A1) * 100	Energy used with Adhoc Requests – Energy used with DWA prediction	(A4/A3) *100
Uid= 148	212.36	69.41	-4.50	93.5167	73.91	34.80
Uid =169	616.56	269.60	-9.03	96.6505	278.64	45.19
Uid = 412	7246.48	6665.69	-218.56	96.7211	6884.25	95.00

Fig. 7.10. Energy saved per job by using LWR in 2697

8. Whether Any Ph.Ds enrolled / produced :

Yes, The Co-investigator of this project (Dr.P.V.Sudha) was awarded Ph.D for her work, which included part of this work.

9. Publications resulting from the project :

The following Papers were Presented as a result of the work done in this Project -

1. " Compendium of load prediction models and approaches " in National Conference on Emerging and Innovative Trends in Computer Science (NCEITCS- 2014) , April 2014 , Hyderabad.
2. "Characterization of Elasticity in Clouds with promise of SLAs " in Fourth International Conference on Advances in Computing and Communications. (ICACC- 2014) ,27- 29 Aug. 2014, Kochi .
3. " **Energy Saving in Cloud by using Enhanced Instance Based Learning (EIBL) for Resource Prediction",** Accepted to be published as a chapter in **Springer - Sustainable Cloud and Energy Services, 2017** , Editor - Wilson Rivera Gallego , Professor of Computer Science and Engineering ,University of Puerto Rico at Mayaguez (UPRM)
4. **Thesis Titled " Prediction of Resource Requirements in Google Compute Cloud " was accepted. Co- investigator of this project (Dr.P.V.Sudha) was awarded Ph.D in 2016 . Part of her work is contribution to this Project.**

10 . References :

[1]C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch, "Heterogeneity and dynamicity of clouds at scale: Google trace analysis," in *Proceedings of the Third ACM Symposium on Cloud Computing*. ACM, 2012, p. 7.

[2]W. Smith, "Prediction services for distributed computing," in *Par- allel and Distributed Processing Symposium, 2007. IPDPS 2007. IEEE International*. IEEE, 2007, pp. 1–10.

[3]P. Garraghan, P. Townend, and J. Xu, "An analysis of the server characteristics and resource utilization in google cloud," in *Cloud Engineering (IC2E), 2013 IEEE International Conference on*. IEEE, 2013, pp. 124–131.

[4]J. Hamilton, "Cooperative expendable micro-slice servers (cems): low cost, low power servers for internet-scale services," in *Confer- ence on Innovative Data Systems Research (CIDRâA˘Z´09)(January 2009)*, 2009.

[5]M. Hogan, F. Liu, A. Sokol, and J. Tong, "N cloud computing standards roadmap," *NIST Special Publication*, vol. 35, 2011.

[6]Z. Dong,W. Zhuang, and R. Rojas-Cessa, "Energy-aware scheduling schemes for cloud data centers on google trace data," in *Green Communications (OnlineGreencomm), 2014 IEEE Online Conference on*. IEEE, 2014, pp. 1–6.

[7]A. E. C. Cloud, "Web page at http://aws. amazon. com/ec2," *Date of last access: 14th September*, 2010.

[8] R. Bryant, A. Tumanov, O. Irzak, A. Scannell,K. Joshi, M. Hiltunen, A. Lagar-Cavilla, and E. De Lara, "Kaleidoscope: cloud microelasticity via vm state coloring," in *Proceedings of the sixth conference on Computer systems*. ACM, 2011, pp. 273–286.

[9]H. Nguyen, Z. Shen, X. Gu, S. Subbiah, and J. Wilkes, "Agile: Elastic distributed resource scaling for infrastructure as a ser- vice," in *Proc. of the USENIX International Conference on Auto- mated Computing (ICACâA˘Z´13). San Jose, CA*, 2013.

[10]A. Cassandra, "Apache cassandra," 2013. [11]T. Clark, "Rightscale," 2010.

[12]A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *ACM SIG- COMM computer communication review*, vol. 39, no. 4. ACM, 2009, pp. 123–134.

[13]A. Ali-Eldin, M. Kihl, J. Tordsson, and E. Elmroth, "Efficient provisioning of bursty scientific workloads on the cloud using adaptive elasticity control," in *Proceedings of the 3rd workshop on Scientific Cloud Computing Date*. ACM, 2012, pp. 31–40.

[14]A. Ali-Eldin, J. Tordsson, E. Elmroth, and M. Kihl, "Workload classification for efficient autoscaling of cloud resources," Tech- nical Report, 2005.[Online]. Available: h ttp://www8. cs. umu. se/research/uminf/reports/2013/013/part1. pdf, Tech. Rep., 2013.

[15]I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and grid computing 360-degree compared," in *Grid Computing Envi- ronments Workshop, 2008. GCE'08*. Ieee, 2008, pp. 1–10.

[16]L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: towards a cloud definition," *ACM SIG- COMM Computer Communication Review*, vol. 39, no. 1, pp. 50– 55, 2008.

[17]M. A Vouk, "Cloud computing–issues, research and implemen- tations," *CIT. Journal of Computing and Information Technology*, vol. 16, no. 4, pp. 235–246, 2008.

[18]A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, "Above the clouds: A berkeley view of cloud computing," *Dept. Electrical Eng. and Com- put. Sciences, University of California, Berkeley, Rep. UCB/EECS*, vol. 28, p. 13, 2009.

[19]Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state- of-the-art and research challenges," *Journal of internet services and applications*, vol. 1, no. 1, pp. 7–18, 2010.

[20]S. K. Garg, C. S. Yeo, A. Anandasivam, and R. Buyya, "Environment-conscious scheduling of hpc applications on dis- tributed cloud-oriented data centers," *Journal of Parallel and Dis- tributed Computing*, vol. 71, no. 6, pp. 732–749, 2011.

[21]D. Warneke and O. Kao, "Exploiting dynamic resource allocation for efficient parallel data processing in the cloud," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 6, pp. 985–997, 2011.

[22]L. Wu, S. K. Garg, and R. Buyya, "Sla-based resource allocation for software as a service provider (saas) in cloud computing envi- ronments," in *Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on*. IEEE, 2011, pp. 195–204.

[23]B. Addis, D. Ardagna, B. Panicucci, and L. Zhang, "Autonomic management of cloud service centers with availability guaran- tees," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*. IEEE, 2010, pp. 220–227.

[24]A. Beloglazov and R. Buyya, "Energy efficient resource manage- ment in virtualized cloud data centers," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. IEEE Computer Society, 2010, pp. 826– 831.

[25]S. Ferretti, V. Ghini, F. Panzieri, M. Pellegrini, and E. Turrini, "Qos–aware clouds," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*. IEEE, 2010, pp. 321–328.

[26]Y. C. Lee, C. Wang, A. Y. Zomaya, and B. B. Zhou, "Profit-driven service request scheduling in clouds," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. IEEE Computer Society, 2010, pp. 15–24.

[27]F. Teng and F. Magoules, "Resource pricing and equilibrium al- location policy in cloud computing," in *Computer and Informa- tion Technology (CIT), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 195–202.

[28]Y. O. Yazir, C. Matthews, R. Farahbod, S. Neville, A. Guitouni, S. Ganti, and Y. Coady, "Dynamic resource allocation in com- puting clouds using distributed multiple criteria decision anal- ysis," in *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*. Ieee, 2010, pp. 91–98.

[29]S. Liu, G. Quan, and S. Ren, "On-line scheduling of real-time services for cloud computing," in *Services (SERVICES-1), 2010 6th World Congress on*. IEEE, 2010, pp. 459–464.

[30]M. Mihailescu and Y. M. Teo, "Dynamic resource pricing on fed- erated clouds," in *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on*. IEEE, 2010, pp. 513–517.

[31]H. N. Van, F. D. Tran, and J.-M. Menaud, "Sla-aware virtual re- source management for cloud infrastructures," in *Computer and Information Technology, 2009. CIT'09. Ninth IEEE International Conference on*, vol. 1. IEEE, 2009, pp. 357–362.

[32]S. Chaisiri, B.-S. Lee, and D. Niyato, "Optimal virtual machine placement across multiple cloud providers," in *Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific*. IEEE, 2009, pp. 103–110.

[33]H. N. Van, F. D. Tran, and J.-M. Menaud, "Performance and power management for cloud infrastructures," in *Cloud Comput- ing (CLOUD), 2010 IEEE 3rd International Conference on*. IEEE, 2010, pp. 329–336.

[34]W. Wang, H. Chen, and X. Chen, "An availability aware virtual machine placement approach for dynamic scaling of cloud appli- cations," in *Ubiquitous Intelligence & Computing and 9th Interna- tional Conference on Autonomic & Trusted Computing (UIC/ATC), 2012 9th International Conference on*. IEEE, 2012, pp. 509–516.

[35]T. Redkar and T. Guidici, *Windows Azure Platform*. Springer, 2009.

[36]A. Singhai, S. Lim, and S. R. Radia, "The scalr framework for internet services," in *Proceedings of the 28th Fault-Tolerant Com- puting Symposium (FTCS-28),* 1998.

[37]M. Mao and M. Humphrey, "Autoscaling to minimize cost and meet application deadlines in cloud workflows," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2011, p. 49.

[38]N. Roy, A. Dubey, and A. Gokhale, "Efficient autoscaling in the cloud using predictive models for workload forecasting," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*. IEEE, 2011, pp. 500–507.

[39]Z. Shen, S. Subbiah, X. Gu, and J. Wilkes, "Cloudscale: elastic resource scaling for multi-tenant cloud systems," in *Proceedings of the 2nd ACM Symposium on Cloud Computing*. ACM, 2011, p. 5.

[40]D. M. Quan, R. Basmadjian, H. De Meer, R. Lent, T. Mahmoodi, D. Sannelli, F. Mezza, L. Telesca, and C. Dupont, "Energy effi- cient resource allocation strategy for cloud data centres," in *Com- puter and Information Sciences II*. Springer, 2012, pp. 133–141.

[41]I. Rodero, J. Jaramillo, A. Quiroz, M. Parashar, F. Guim, and

S. Poole, "Energy-efficient application-aware online provisioning for virtualized clouds and data centers," in *Green Computing Con- ference, 2010 International*. IEEE, 2010, pp. 31–45.

[42]H. S. Abdelsalam, K. Maly, R. Mukkamala, M. Zubair, and

D. Kaminsky, "Analysis of energy efficiency in clouds," in *Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, 2009. COMPUTATIONWORLD'09. Computation World:*. IEEE, 2009, pp. 416–421.

[43]A. J. Younge, G. Von Laszewski, L. Wang, S. Lopez-Alarcon, and

W. Carithers, "Efficient resource management for cloud comput- ing environments," in *Green Computing Conference, 2010 Inter- national*. IEEE, 2010, pp. 357–364.

[44]P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being green," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 211–222, 2012.

[45]D. Wang, C. Ren, A. Sivasubramaniam, B. Urgaonkar, and Fathy, "Energy storage in datacenters: what, where, and how much?" in *ACM SIGMETRICS Performance Evaluation Review*, vol. 40, no. 1. ACM, 2012, pp. 187–198.

[46]M. Kurpicz, A. Sobe, and P. Felber, "Using power measurements as a basis for workload placement in heterogeneous multi-cloud environments," in *Proceedings of the 2nd International Workshop on CrossCloud Systems*. ACM, 2014, p. 6.

[47]F. Chen, J. Grundy, J.-G. Schneider, Y. Yang, and Q. He, "Au- tomated analysis of performance and energy consumption for cloud applications," in *Proceedings of the 5th ACM/SPEC interna- tional conference on Performance engineering*. ACM, 2014, pp. 39–50.

[48]C. You and K. Chandra, "Time series models for internet data traffic," in *Local Computer Networks, 1999. LCN'99. Conference on*. IEEE, 1999, pp. 164–171.

[49]P. J. Brockwell, *Introduction to time series and forecasting*. Taylor & Francis, 2002, vol. 1.

[50]C. Chatfield, *The analysis of time series: an introduction*.CRC press, 2013.

[51]P. A. Dinda and D. R. O'Hallaron, "Host load prediction using linear models," *Cluster Computing*, vol. 3, no. 4, pp. 265–280, 2000.

[52]J. Tikka, A. Lendasse, and J. Hollmén, "Analysis of fast input se- lection: Application in time series prediction," in *Artificial Neural Networks–ICANN 2006*. Springer, 2006, pp. 161–170.

[53]D. Gmach, J. Rolia, L. Cherkasova, and A. Kemper, "Capacity management and demand prediction for next generation data centers," in *Web Services, 2007. ICWS 2007. IEEE International Conference on*. IEEE, 2007, pp. 43–50.

[54]E. S. Buneci and D. A. Reed, "Analysis of application heartbeats: Learning structural and temporal features in time series data for identification of performance problems," in *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*. IEEE Press, 2008, p. 52.

[55] W. Fang, Z. Lu, J. Wu, and Z. Cao, "Rpps: A novel resource prediction and provisioning scheme in cloud data center," in *Ser- vices Computing (SCC), 2012 IEEE Ninth International Conference on*. IEEE, 2012, pp. 609–616.

[56] J. Huang, C. Li, and J. Yu, "Resource prediction based on double exponential smoothing in cloud computing," in *Consumer Elec- tronics, Communications and Networks (CECNet), 2012 2nd Inter- national Conference on*. IEEE, 2012, pp. 2056–2060.

[57] J. Geweke and C. Whiteman, "Bayesian forecasting," *Handbook of economic forecasting*, vol. 1, pp. 3–80, 2006.

[58]S. Di, D. Kondo, and W. Cirne, "Host load prediction in a google compute cloud with a bayesian model," in *Proceedings of the In- ternational Conference on High Performance Computing, Network- ing, Storage and Analysis*. IEEE Computer Society Press, 2012, p. 21.

[59]S. Kundu, R. Rangaswami, A. Gulati, M. Zhao, and K. Dutta, "Modeling virtualized applications using machine learning tech- niques," in *ACM SIGPLAN Notices*, vol. 47, no. 7. ACM, 2012, pp. 3–14.

[60]P. Lama and X. Zhou, "Autonomic provisioning with self-adaptive neural fuzzy control for end-to-end delay guarantee," in *Mod- eling, Analysis & Simulation of Computer and Telecommunica- tion Systems (MASCOTS), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 151–160.

[61]L. Aranildo Rodrigues, P. S. de Mattos Neto, and T. Ferreira, "A prime step in the time series forecasting with hybrid methods: The fitness function choice," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. IEEE, 2009, pp. 2703– 2710.

[62]K. Mohammadi, H. Eslami, S. D. Dardashti, *et al.*, "Comparison of regression, arima and ann models for reservoir inflow forecast- ing using snowmelt equivalent (a case study of karaj)," *J. Agric. Sci. Technol*, vol. 7, pp. 17–30, 2005.

[63]Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Network and Service Management (CNSM), 2010 International Conference on*. IEEE, 2010, pp. 9–16.

[64]Y. Jiang, C.-s. Perng, T. Li, and R. Chang, "Asap: A self-adaptive prediction system for instant cloud resource demand provision- ing," in *Data Mining (ICDM), 2011 IEEE 11th International Confer- ence on*. IEEE, 2011, pp. 1104–1109.

[65]T. Miu and P. Missier, *Predicting the Execution Time of Workflow Blocks Based on Their Input Features*. Computing Science, New- castle University, 2013.

[66]C.-Y. Lin, Y.-A. Chen, Y.-C. Tseng, L.-C. Wang, *et al.*, "A flexible analysis and prediction framework on resource usage in public clouds." in *CloudCom*, 2012, pp. 309–316.

[67]N. R. Herbst, N. Huber, S. Kounev, and E. Amrehn, "Self-adaptive workload classification and forecasting for proactive resource provisioning," *Concurrency and Computation: Practice and Expe- rience*, vol. 26, no. 12, pp. 2053–2078, 2014.

[68]G. Wang, A. Khasymski, K. Krish, and A. R. Butt, "Towards improving mapreduce task scheduling using online simulation based predictions," in *Parallel and Distributed Systems (ICPADS), 2013 International Conference on*. IEEE, 2013, pp. 299–306.

[69]D. T. Pham, S. S. Dimov, and C. Nguyen, "Selection of k in k- means clustering," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 219, no. 1, pp. 103–119, 2005.

[70]C.-F. Lai, J.-H. Chang, C.-C. Hu, Y.-M. Huang, and H.-C. Chao, "Cprs: A cloud-based program recommendation system for dig- ital tv platforms," *Future Generation Computer Systems*, vol. 27, no. 6, pp. 823–835, 2011.

[71]P. W. Buana and I. K. G. Darma, "Combination of k-nearest neighbor and k-means based on term re-weighting for classify in- donesian news," *International Journal of Computer Applications*, vol. 50, no. 11, 2012.

[72]G. Batista and D. F. Silva, "How k-nearest neighbor parameters affect its performance," in *Argentine Symposium on Artificial In- telligence*, 2009, pp. 1–12.

[73]F. Farahnakian, A. Ashraf, P. Liljeberg, T. Pahikkala, J. Plosila, Porres, and H. Tenhunen, "Energy-aware dynamic vm consol- idation in cloud data centers using ant colony system," in *Cloud Computing (CLOUD), 2014 IEEE 7th International Conference on*. IEEE, 2014, pp. 104–111.

[74]W. S. Cleveland and S. J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596– 610, 1988.

[75]N. H. Kapadia, J. A. Fortes, and C. E. Brodley, "Predictive application-performance modeling in a computational grid envi- ronment," in *High Performance Distributed Computing, 1999. Pro- ceedings. The Eighth International Symposium on*. IEEE, 1999, pp. 47–54.

[76]C. G. Atkeson, A. W. Moore, and S. Schaal, "Locally weighted learning for control," in *Lazy learning*. Springer, 1997, pp. 75– 113.

[77]M. Kayri and G. Zırhlıo "Kernel smoothing function and choosing bandwidth for non-parametric regression methods," *Ozean Journal of Applied Sciences*, vol. 2, no. 1, pp. 49–54, 2009.

[78]B. J. Barnes, B. Rountree, D. K. Lowenthal, J. Reeves, B. De Supinski, and M. Schulz, "A regression-based approach to scalability prediction," in *Proceedings of the 22nd annual interna- tional conference on Supercomputing*. ACM, 2008, pp. 368–377.

[79]Q. Zhang, L. Cherkasova, and E. Smirni, "A regression-based analytic model for dynamic resource provisioning of multi-tier applications," in *Autonomic Computing, 2007. ICAC'07. Fourth In- ternational Conference on*. IEEE, 2007, pp. 27–27.

[80]H. A. Guvenir and I. Uysal, "Regression on feature projections," *Knowledge-Based Systems*, vol. 13, no. 4, pp. 207–214, 2000.

[81]A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, "To- wards characterizing cloud backend workloads: insights from google compute clusters," *ACM SIGMETRICS Performance Evalu- ation Review*, vol. 37, no. 4, pp. 34–41, 2010.

[82]S. Aggarwal, S. Phadke, and M. Bhandarkar, "Characterization of hadoop jobs using unsupervised learning," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second Interna- tional Conference on*. IEEE, 2010, pp. 748–753.

[83]J. Tan, P. Dube, X. Meng, and L. Zhang, "Exploiting resource us- age patterns for better utilization prediction," in *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on*. IEEE, 2011, pp. 14–19.

[84]Q. Zhang, J. L. Hellerstein, and R. Boutaba, "Characterizing task usage shapes in google's compute clusters," in *Large Scale Distributed Systems and Middleware Workshop (LADIs´11)*, 2011.

[85]B. Sharma, V. Chudnovsky, J. L. Hellerstein, R. Rifaat, and C. R. Das, "Modeling and synthesizing task placement constraints in Google compute clusters," in *Proceedings of the 2nd ACM Symposium on Cloud Computing*. ACM, 2011, p. 3.

[86]S. Di, D. Kondo, and W. Cirne, "Characterization and comparison of cloud versus grid workloads," in *Cluster Computing (CLUSTER), 2012 IEEE International Conference on*. IEEE, 2012, pp. 230– 238.

[87]I. S. Moreno, P. Garraghan, P. Townend, and J. Xu, "An approach for characterizing workloads in Google cloud to derive realistic re- source utilization models," in *Service Oriented System Engineering (SOSE), 2013 IEEE 7th International Symposium on*. IEEE, 2013, pp. 49–60.

[88]M. Amoretti, A. L. Lafuente, and S. Sebastio, "A cooperative approach for distributed task execution in autonomic clouds," in *Parallel, Distributed and Network-Based Processing (PDP), 2013 21st Euromicro International Conference on*. IEEE, 2013, pp. 274–281.

[89]J. L. Hellerstein, W. Cirne, and J. Wilkes, "Google cluster data," *Google research blog, Jan*, 2010.

[90]Y. Jiang, C.s. Perng, T. Li, and R. Chang, "Self-adaptive cloud capacity planning," in *Services Computing (SCC), 2012 IEEE Ninth International Conference on*. IEEE, 2012, pp. 73–80.

[91]S. Kavulya, J. Tan, R. Gandhi, and P. Narasimhan, "An analysis of traces from a production mapreduce cluster," in *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on*. IEEE, 2010, pp. 94–103.

[92]A. Beloglazov and R. Buyya, "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 eScience*. ACM, 2010, p. 4.

[93]X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ACM SIGARCH Computer Archi- tecture News*, vol. 35, no. 2.ACM, 2007, pp. 13–23.

[94]J. Zheng, T. S. E. Ng, and K. Sripanidkulchai, "Workload-aware live storage migration for clouds," in *ACM SIGPLAN Notices*, vol. 46, no. 7. ACM, 2011, pp. 133–144.

[95]S. Daniel and M. Kwon, "Prediction-based virtual instance migration for balanced workload in the cloud datacenters," 2011.

[96]A. Abdulmohson, S. Pelluri, and R. Sirandas, "Energy efficient load balancing of virtual machines in cloud environments," 2015.

[97]C. Lively, X. Wu, V. Taylor, S. Moore, H.-C. Chang, and . Cameron, "Energy and performance characteristics of different parallel implementations of scientific applications on multi- core systems," *International Journal of High Performance Computing Applications*, vol. 25, no. 3, pp. 342–350, 2011.

[98]D. Kliazovich, P. Bouvry, and S. U. Khan, "Simulation and performance analysis of data intensive and workload intensive cloud computing data centers," in *Optical Interconnects for Future Data Center Networks*. Springer, 2013, pp. 47–63.

[99]S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in *Proceedings of the 2008 conference on Power aware computing and systems*, vol. 10. San Diego, California, 2008.

[100]D. Kusic, J. O. Kephart, J. E. Hanson, N. Kandasamy, and G. Jiang, "Power and performance management of virtualized computing environments via lookahead control," *Cluster computing*, vol. 12, no. 1, pp. 1–15, 2009.