Abstract: Cloud computing is an infrastructure for running enterprise and Web applications in cost effective manner. However, the growing needs of Cloud have increased the energy consumption of data centers, which has become a major issue. High energy consumption not only translates to high operational cost but also badly affects the environment. So the design of the cloud is such that must be power efficient. The main goal of this paper is to improve the utilization of computing resources and reduce energy consumption under workload independent quality of service constraints. This paper also describes the migration of application towards the cloud and discusses various elements of Cloud which contribute to the total energy consumption.

Keywords: Cloud computing, energy efficiency, scheduling, cluster, migration.

I. INTRODUCTION

Cloud Computing is a new consumption and virtualization model for the high cost computing infrastructures and is a web based IT solution. Cloud provides suitable, on-demand service, network access, resource sharing and measured service in highly customizable manner with minimal management effort. The application of low-cost computing devices, high-performance network resources, huge storage capacity, semantic web technology, etc., have helped in the swift growth of cloud technology. A cloud infrastructure generally encapsulates all those existing technologies in a web service based model to offer improved scalability and on demand availability. The rapid deployment model, low start up investment, pay-as-you-go scheme, multi-tenant sharing of resources are all added attributes of cloud technology due to which major industries tend to virtualization for their enterprise applications.

Cloud applications are deployed in remote data centers (DCs) where high capacity servers and storage systems are located. A fast growth of demand for cloud based services results into establishment of enormous data centers consuming high amount of electrical power. Energy efficient model is required for complete infrastructure to reduce functional costs while maintaining vital Quality of Service (QoS). Energy optimization can be achieved by combining resources as per the current utilization, efficient virtual network topologies and thermal status of computing hardware and nodes. On the other hand, the primary motivation of cloud computing is related to its flexibility of resources.

In this paper we plan to consolidate all the aspects of energy efficient infrastructure model for cloud data centers while considering performance bottlenecks for the same and also migration of applications towards the cloud.

A) Problem Definition

This paper mainly tackle the problems in relation to energy efficiency in cloud computing. Major challenges in this research are the development of a system which consumes less energy and less cost. In particular, the following research problems are investigated:

- Develop a cloud infrastructure in energy efficient manner.
- Built a concept that supports the organization to take the proper decision in cloud migration.
- Analysis the energy consumption.

II. CLOUD COMPUTING

Cloud computing is a paradigm which enables the outsourcing of all IT needs such as storage, computation and software such as office and ERP, through large Internet. The literary meaning of “Cloud computing” can be “computing achieved using collection of networked resources, which are offered on subscription”. Internet is a fundamental medium through which these Cloud services are made accessible and delivered to end user. The growing popularity of Cloud computing has led to several proposal defining its characteristics. Some of the definitions given by many well-known scientists and organizations include:

A) Buyya et al. define the Cloud computing in terms of its utility to end user: “A Cloud is a market-oriented distributed computing system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers.”

B) National Institute of Standards and Technology (NIST) [7] defines Cloud computing as follows: “Cloud
computing is a model for enabling 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 management effort or service provider interaction. This Cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.”

The characteristics of Clouds include on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. The available service models are classified as SaaS (Software-as-a-Service), PaaS (Platform-as-a-Service), and IaaS (Infrastructure-as-a-Service). The deployment models is categorised into public, private, community, and hybrid Clouds.

III. THE FIRST STEP: ENERGY EFFICIENCY IN CLOUD INFRASTRUCTURES

Building an energy efficient cloud model does not indicate only energy efficient host machines. Other existing components of a complete cloud infrastructure should also be considered for energy aware applications. Several research works have been carried out to build energy efficient cloud components individually. In this section we will investigate the areas of a typical cloud setup that are responsible for considerable amount of power dissipation and we will consolidate the possible approaches to fix the issues considering energy consumption as a part of the cost functions to be applied.

A) Energy Efficient Hardware

One of the best approaches to reduce the power consumption at data centre and virtual machine level is usage of energy efficient hardwares at host side. New electronics materials like solid-state drives are more power efficient than common hard disk drives but that are costly. Some existing mechanisms for saving energy at hardware level are given below:

- SpeedStep®: This is Intel’s wireless technology to adjust CPU power dynamically based upon the performance demand. It works in five usage modes: standby mode, voice communication, multimedia, data communication, and multimedia and data communication.
- PowerNow™: This is AMD’s power saving technology which can manage power consumption instantly, on-the-fly by controlling voltage and frequency independently. It can be operated in three modes namely High Performance mode, Power saver mode and automatic mode.
- Cool’n’Quit™: AMD’s Cool’n’Quit™ technology controls the system fan, voltage and clock speed of the processor’s cores based on the system temperature.

All these technologies are able to sense lack of machine interaction and then different hardware parts can incrementally be hibernated or put in sleep mode to save energy.

B) Energy Efficient Resource Scheduling

Several research works have been carried out on energy efficient resource scheduling in virtual machines and grid systems. Many aspects such as: power consumption, reliability, response time etc. need to be considered while designing VM schedulers.

1) Real time tasks scheduling in multiprocessor systems: Real time scheduling of tasks in a multiprocessor host machine has been a major issue to be considered. Energy aware, real-time task scheduling method in multiprocessor system supports Dynamic Voltage Scaling (DVS). It is a polynomial time heuristic approach for building probabilistic model of load balancing issue. Task scheduling can be of two types: i) partitioned scheduling – where assignment of each individual task to a particular processor is fixed and ii) non-partitioned (dynamic) scheduling – where a long term global scheduler assigns the jobs to processors on-the-fly from a ready queue. In a multiprocessor system, there is a set of pre-emptive, independent, real time tasks denoted as \( T = \{ T_1, T_2, \ldots, T_n \} \) and a set of processor cores denoted as \( P = \{ P_1, P_2, \ldots, P_m \} \) where each core has a finite range of discrete frequencies \( F = \{ f_1, f_2, \ldots, f_k \} \).

2) Memory-aware Scheduling in multiprocessor systems: Optimization of scheduling algorithm helps to reduce the energy consumption by memory and also affects the efficiency of frequency scaling. In present multicore systems, cores on the chip share resources...
such as caches, DRAM etc. Tasks running on one core may harmfully affect the performance of tasks on other cores and hence it may even maliciously create a Denial of Service (DoS) attack on the same chip.

C) Energy efficient Network Infrastructure in cloud

Network in a cloud environment can be of two types - wireless network and wired network. According to ICT, energy estimates the radio access network consumes a major part of the total energy in an infrastructure and the cost incurred on energy consumption is sometimes comparable with the total cost spent on personnel employed for network operations and maintenance. In order to achieve energy savings, use GAF and CEC protocols. These protocols identify redundant nodes and turn off them to conserve energy.

1) Geographic Adaptive Fidelity Protocol: In GAF protocol, equivalent nodes are found out by using their geographical information and then their radios are turned off which saves energy. However for communication between a pair of nodes, nodes which are equal may not be equal for communication between a different pair. This problem is addresses by dividing the whole network into virtual grids which have the property that all nodes in adjacent grids can communicate with each other. All nodes within a single grid are equivalent. Also the nodes in GAF protocol always switch among one of the three states – sleeping, discovery and active. Initially a node is in discovery state with its radio turned on and it exchanges messages with its neighbors. A node in active state and discovery state can switch to sleeping state whenever it finds an equivalent node which can perform routing.

![Figure 3: State transition in GAF](image)

2) Cluster Based Energy Conservation Protocol: One of the disadvantages with GAF protocol is that it needs global location information which may not be available every time. Also it is very conservative because it guesses its connectivity instead of directly measuring it which leads to less energy savings. Another protocol namely Cluster based Energy Conservation overcomes these disadvantages since it is independent of location information and finds network redundancy more accurately thus saving more energy. CEC operates in following steps-

- Determining Network Redundancy: It operates by organizing nodes into clusters which are overlapping. A cluster can be viewed as a circle around the node called cluster head. A cluster head is defined in such a way that it reaches each node in a cluster in just one hop. Another node called gateway node interconnects two or more clusters is a member of multiple clusters and overall network connectivity is ensured by it. Third type of node in a cluster is called ordinary node and is the redundant one because it neither the gateway node nor the cluster head.

- Distributed Cluster formation: A cluster selects a cluster-head and broadcast node by broadcasting discovery messages to its neighbors. A discovery message is a combination of node ID, its cluster ID and its estimated lifetime. A node having the longest lifetime selects itself as the cluster head and this information is obtained by exchanging discovery messages. The node id of the cluster head becomes the cluster id of each node in a cluster. A primary gateway node is one that can hear multiple cluster heads and the one hearing the gateway node and cluster head is secondary gateway node. Some of the redundant gateway nodes among multiple gateway nodes between two clusters are suppressed by CEC to conserve energy. These redundant gateway nodes are selected according to certain rules. Primary gateway nodes have higher priority and are preferred over secondary because only one primary node can connect adjacent clusters. Finally redundant nodes are powered off after selecting cluster-heads and gateway-nodes thus conserving energy.

![Figure 4: Cluster formation in CEC protocol](image)

Energy savings in wired networks is not given as much importance as compared to energy savings in wireless networks. In fact, concepts applied for energy conservation in wireless networks cannot directly be applied in wired networks such as turning the nodes power off when not required to save energy because of high volumes and rate of data traffic and further nodes need to satisfy quality of constraint requirements. Also in a wired network, 60% of the energy consumed may be due to peripheral devices such as link drivers.

IV. SECOND STEP: MIGRATION OF THE APPLICATION TO THE CLOUD

Application migration is the process of transitioning all or parts of an enterprise data, applications and services from on-site premises and redeploying them on a new platform and infrastructure in the cloud. The migration process includes the staging of the new environment before the actual cutover for the application and requires
coordination between the IT teams in the time of migration.

A) Will all the application move to the cloud?

The answer is no. Some applications can be considered as good candidates to move fully or partially to the cloud platform, while other applications do not fit the cloud. To identify which applications can move to the cloud, applications are decomposed into several attributes form technical and business factors. The cloud is also decomposed into attributes to help make decision to run the application in the cloud or not.

1) Cloud computing is a Fit

- Database Intensive Applications: It fits the cloud as the cloud computing and database are compatible. However, the cloud is more suitable for storing large volumes of unstructured data. On the other hand, when the data is accessed frequently in huge number of transactions in applications such as banking systems, then it can struggle the cloud. Such applications are governed by hardware limitations inside all IT environments, and thus there are no or little benefits from sending them to the cloud.

- Network Intensive Applications: It is a good fit, but it is necessary to have an access to the fastest and highest quality network services in the cloud. Attributes such as network capacity, latency, redundancy and routing flexibility have to be taken into account when searching for the proper cloud provider.

- Virtualized Applications: It is an excellent fit. The cloud is completely compatible with such applications. The best way to ensure the cloud success is to optimize the applications in terms of CPU, storage, interface, and network performance.

2) Cloud Computing is not a Fit: Practically, the opposite characteristics for the applications or the classifications of applications mentioned in the previous section are not fit for cloud computing. Here more detailed information will be added about the applications and the characteristics of the applications which are not recommended to move to the cloud.

- Highly Secured Applications: when a high level of security is required, or the application has very sensitive data, then moving application to the cloud is not highly recommended. When company moves its IT service on the cloud, they are no longer in control of the location of data. Thus, the organization has to be careful when investigating the security level/controls that the service provider ensures.

- When the control is very critical on the business: some components in the applications are very critical to the business. If such components will not be moved to 100% reliable cloud platform, then it could be endangering.

B) Attributes in the Migration Decision

The following shows some of these attributes and their relations to cloud computing based on some findings extracted mainly from the advantages and disadvantages of cloud, or in others word, what the cloud likes and what does not like.

1) Online Data Access: this attribute affect the decision when the company wants to use the cloud storage services. If the application requires online access for data, then cloud would be an excellent choice as the company does not need to keep any data on-site, and thus it does not need to move huge bulk of data to their local servers.

2) Offline Data Access: on the contrast to online access data, when an organization requires pure offline access of data, then it is better to keep data locally rather than put it on the cloud.

3) Structured Data: cloud supports structured data by services that provide database-like functionality. However, when the application is a database intensive with huge amount of transactions, this can cause some reliability issues. Structured data requires relational database such as SQL databases; SQL databases are difficult to scale, meaning that they are not natively suited to a cloud environment.

4) Unstructured Data: it is more supported by cloud environment. An example of such data is a file system that stores media files. Unstructured data is non-relational or NoSQL database; it is built to service heavy read and write operations and it is able to scale up and down easily, and thus it is more natively suited to run on the cloud.

5) Authorization/Authentication: the enterprise authentication and authorization framework does not naturally extend into the cloud. The multi-tenancy architecture of the cloud force different mechanisms for authentication and authorization; cloud application services multiple tenants using one server and one application instance. Thus, there will be two kind of identification (authentication): tenant and user identification. Tenant can be identified by the domain name of the email address for example, while the user identification can be done by using special user database for every tenant (company).

6) Replication: cloud is tightly related to replication technology. Replication can be performed for data or computing nodes. When the main replica crashes, one of the other replicas compensate the failed one instantaneously. Moreover, these replicas can be used in the load distribution so the application can handle the increasing demands.

7) Caching: cloud supports caching by adding another storage layer “distributed data cache” makes data access quick.
8) Pooling: Cloud is a shared pool of computing resources such as network, server, storage, application, and service. Pooling is strongly correlated to the cloud environment.

9) Scale Out/Up: application migrated to the cloud can benefit from elastic nature of cloud computing, and therefore, scale-out/scale-up attributes are supported by the cloud with pay-as-you-use purchase model.

C) Migration Decision Calculation Process (Migration Algorithm)

In the preceding section, we have explained the role of each attribute and low level attribute in the migration decision to cloud computing environment, and how it can affect the organizations’ decisions when they think to move their applications to the cloud. Every organization has its own combination of these attributes according to how these attributes are more critical for their businesses. For example, some organizations concerns in scalability more than affordability, while others are aware of the security and availability issues. As mentioned before, the first step in the concept is to tell the user (organization) how good or not its decision to move the application to the cloud. Thus, the system takes this preferable combination from different kind of attributes and makes some calculations (Migration algorithm) on them, and returns to the user the score of this combination and the recommended decision that organization should take. The following steps show in details the process to perform these calculations (Migration Algorithm).

1) Every attribute has been given a certain value (weight) represents the contribution of this attribute in cloud computing. In other words, how much the cloud computing supports the attribute. For example, online data access has been given a positive value as it is recommended by the cloud, while offline attribute has been given a negative value as it is not recommended moving the application that accesses its data offline to the cloud.

2) The weights mentioned in the first point are given for every application attribute. The idea behind these weights is to measure the impact for the attribute in the application migration decision as they will be used in the equation in the point 6; the attribute has a stronger relation with cloud will have higher weight and indeed will be weighted more in the equation. Not all weights are given positive values, some attributes are given negative value and these attributes can decrease the application score in the equation if they have been selected by the user, and consequently, reduce the chance to migrate the application to the cloud.

3) The organization can select any number or combination from these attributes even though that some attributes can have conflict contribution in the cloud decision. For instance, user can select both online and offline data access or structured and unstructured data storage which is a conflicted attributes combination.

4) To solve the issue of selecting some conflicts in the selected attributes, we added levels (importance levels) for every selected attribute such as low, high and critical. For example, if the user selects offline and online data attributes, online attribute can be more important for the application than offline, then the user is able to choose “High” or “Critical” levels for online and low level for offline.

5) In addition to the weights given to the attributes as numerical values, the levels (low, high, critical) are given numerical values as well. Low, high, and critical are assigned to 1, 2, and 3 respectively. Both weights and levels are used to calculate the application score which controls the migration decision.

6) From the selected combination from the attributes and their levels, the business logic (migration algorithm) calculates the following:

- The lowest value in the score scale (the lowest application score): it is the lowest score can be obtained from any selected combination the user can select from the available attributes. It is calculated as if the user selects the most negative combination of attributes and levels. To clarify, user only selects the attributes which have negative weights. Moreover, user selects “Critical” as level for every negative attribute. Thus, any application score can be obtained later will not be less than this value. This lowest value represents will be the lowest point in the “ruler” that will measure the application score for the migration decision.

- The highest value in the score scale (the highest application score): it is the highest score can be obtained from any selected combination the user can select from the available attributes. It is calculated as if the user selects the most positive combination of attributes and levels. To clarify, user only selects the attributes which have positive weights. Moreover, user selects “Critical” as level for every positive attribute. Thus, any application score can be obtained later will not be higher than this value. This lowest value represents will be the highest point in the “ruler” that will measure the application score for the migration decision.

- The real score value: it is the real score of the selected combination, and depending on this value the decision is taken to move the application to the cloud or not. The value is calculated according to the following formula:

   Application Score = ΣAw×Lv

   Where Aw = the weight of the attribute.
   Lv = the value of the level

7) From the lowest the highest application scores calculated from the previous points, we define the scale “ruler” that can measure the score of the application to move to the cloud or not.

8) To add more flexibility to the score scale, we divided the resulted scale into five periods. Two periods in the
negative side, two periods in the positive side, and one in the middle around \{0\} as follows:

- The middle period starts with a certain value represent percentage from the lowest score. This percentage has been taken as 10%. The middle point ends with a value equal to 10% from the highest score. The idea behind this small period compared to the other four periods that we want to make a period where the user can decide between move the application to the cloud or not.
- The remaining scale is divided equally in both sides.

9) When the application’s score comes in the first period that means that the application is not a cloud fit completely, while if the score comes in the last period that means that the application is strongly recommended to move to the cloud. The middle period gives the user the freedom to choose either to move the application or keep the thing as they are. The other two periods: one comes in the negative area and it does not support the migration decision, while the other one located in the positive area and it recommends the user to move the application to the cloud.

V. ENERGY CONSUMPTION ANALYSIS

1) To achieve energy efficiency at application level, SaaS providers should pay attention in deploying software on right kind of infrastructure which can execute the software most efficiently. For example, a simple task such as listening to music can consume significantly different amounts of energy on a variety of heterogeneous devices. As task have the same purpose on each device, the results show that the implementation of the task and the system upon which it is performed can have a dramatic impact on efficiency.

2) To calculate the amount of energy consumed by data centers, two metrics were established by Green Grid, an international consortium. The metrics are Power Usage Effectiveness (PUE) and Data Centre Infrastructure Efficiency (DCiE) as defined below:

\[
PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}
\]

\[
DCiE = \frac{1}{PUE} = \left(\frac{\text{IT Equipment Power}}{\text{Total Facility Power}}\right) \times 100\%
\]

The IT equipment power is the load delivered to all computing hardware resources, while the total facility power includes other energy facilities, specifically, the energy consumed by everything that supports IT equipment load.

3) In cloud infrastructure, a node refers to general multicore server along with its parallel processing units, network topology, power supply unit and storage capacity. The overall energy consumption of a cloud environment can be classified as follows:

\[
E_{\text{Cloud}} = E_{\text{Node}} + E_{\text{Switch}} + E_{\text{Storage}} + E_{\text{Others}}
\]

Consumption of energy in a cloud environment having \(n\) number of nodes and \(m\) number of switching elements can be expressed as:

\[
m(E_{\text{Chassis}} + E_{\text{Linecards}} + E_{\text{Ports}}) + (E_{\text{NAS Server}} + E_{\text{Storage Controller}} + E_{\text{Disk Array}}) + E_{\text{Others}}.
\]

VI. LITERATURE SURVEY

Arindam Banerjee, Prateek Agrawal and N. Ch. S. N. Iyengar[1] investigate all possible areas in a typical cloud infrastructure that are responsible for substantial amount of energy consumption and address the methodologies by which power utilization can be decreased without compromising Quality of Services (QoS) and overall performance. And also define the scope for further extension of research from the findings.

Anton Beloglazov and Prof. Rajkumar Buyya[2] describes that, in 2010, energy consumption by data centers worldwide was estimated to be between 1.1% and 1.5% of the global electricity use and is expected to grow further. This thesis presents novel techniques, models, algorithms, and software for distributed dynamic consolidation of Virtual Machines (VMs) in Cloud data centers. The goal is to improve the utilization of computing resources and reduce energy consumption under workload independent quality of service constraints. The proposed approach is distributed, scalable, and efficient in managing the energy-performance trade-off.

Abdallah Issa Salama[3] focuses on how to employ the cloud computing technology as an energy efficient solution. The main idea applied in the concept is identifying architectures for cloud applications depending on their inherent properties of cloud computing such as virtualization and the elasticity that can make them green potential, and identifying correlations between these architectures with already identified business process patterns used in green business process design.

Ashkan Gholamhosseinian & Ahmad Khalifeh[4] describe the energy efficient approaches inside data centres from the site and IT infrastructure perspective incorporating. Virtualization as heart of energy efficient Cloud computing that can integrates some technologies like consolidation and resource utilization has been introduced to prepare a background for implementation part. Finally approaches for Cloud computing data centres at operating system and especially data centre level are presented and Green Cloud architecture as the most suitable green approach is described in details. In the experiment segment we modelled and simulated Facebook and studied the behaviour in terms of cost and performance and energy consumption to reach a most appropriate solution.
Saurabh Kumar Garg and Rajkumar Buyya[5] discuss various elements of Clouds which contribute to the total energy consumption and how it is addressed in the literature. We also discuss the implication of these solutions for future research directions to enable green Cloud computing and also explains the role of Cloud users in achieving this goal.

VII. CONCLUSION

In this paper we have investigated the need of power consumption and energy efficiency in cloud computing model. It has been shown that there are few major components of cloud architecture which are responsible for high amount of power dissipation in cloud. The possible ways to meet each sector for designing an energy efficiency model has also been studied. And also study about the migration of application in the cloud i.e. which application is going to migrate to cloud or which is not. Finally we have found the energy consumption analysis.

REFERENCE


[5] Saurabh Kumar Garg and Rajkumar Buyya “Green Cloud computing and Environmental Sustainability” Dept. of Computer Science and Software Engineering, The University of Melbourne, Australia


