IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

A Pragmatic Approach to Optimize Energy Efficient Resource Allocation Technique in Cloud Computing Data Center

Neha Pangotra[1], Meenakshi Sharma[2]

Department of Computer Science and Engineering, SSCET, Pathankot, India[1, 2]

Abstract: Rapid growth of the demand for computational power has led to the creation of large-scale cloud computing data centers.The development of computing systems has always been focused on performance improvements driven by the demand of applications from consumer, scientific and business domains, but the ever increasing energy consumption of computing systems has started to limit further performance growth due to overwhelming energy consumption and carbon dioxide footprints. Hence, the goal of the computer system design has been shifted from performance improvements to power and energy efficiency. Data centers consume enormous amounts of electrical power resulting in high operational costs and carbon dioxide emissions. Moreover, modern Cloud computing environments have to provide high Quality of Service (QoS) for their customers resulting in the necessity to deal with power-performance trade-off. Reducing carbon emission by cloud computing data centers has emerged as one the dominant research topics both in industry and academia. The foremost objective of cloud service providers is to have a cost effective and energy efficient solution for allocating virtualized ICT resources to end-users' application while meeting the QoS (Quality of Service) level as per SLA (Service Level Agreement).This paper presents a hybrid energy efficient resource allocation technique which combines predictive with reactive allocation techniques and accomplishes substantial improvements in: (a) meeting SLAs, (b) conserving energy, and (c)meeting static and dynamic resource allocation. In this paper we propose energy-aware allocation heuristics provision data center resources to client applications in a way that utilises the capability of VMs live migration to reallocate resources dynamically and improves energy efficiency of the data center, while delivering the negotiated Quality of Service (QoS). The basic idea is to use a heuristic that is consolidating and rearranging the allocation of resources in an energy efficient manner.

Keywords: Cloud Computing, Energy Efficiency, Resource Allocation, Virtualization, virtual machines, Data Center.

I. INTRODUCTION

Cloud computing is today's most emphasized Information and Communications Technology (ICT) paradigm that is directly or indirectly used by almost every online user. However, such great significance comes with the support of a great infrastructure that includes large data centers comprising thousands of server units and other supporting equipment. Their share in power consumption generates between 1.1% and 1.5% of the total electricity use worldwide and is projected to rise even more. Such alarming numbers demand rethinking the energy efficiency of such infrastructures. Cloud computing allows business customers to scale up and down their resource usage based on needs. Many of the touted gains in the cloud model come from resource multiplexing through virtualization technology. In this paper, we present a system that uses virtualization technology to allocate data center resources predictably as well as reactively based on application demands and support energy efficient cloud computing by optimizing the number of servers and memory in use. We use the concept of minimization of CPU utilization and RAM to measure the overutilization and underutilization of servers to achieve better energy efficient and resource allocation technique by providing Quality of Service while meeting desired SLA. This helps to improve the overall utilization of resources in cloud data centers. We develop a set of heuristics that prevent overload in the system effectively while saving energy used. This paper leverages virtualization technology which provides mechanism for mapping virtual machines (VMs) to physical resources. Hardware virtualization hides the underlying computing system to present an abstract computing platform by using a hypervisor. In data centres, the number of physical machines can be reduced using virtualization by consolidating VMs onto shared servers which helps improve the efficiency of IT systems. It allows multiple virtual machines to be run on a single physical machine in order to provide more capability and increase the utilization level of the hardware. The main instrument that we leverage is live migration of VMs which is achieved using CloudSim. The ability to migrate VMs between physical hosts with low overhead gives flexibility to a resource provider as VMs can be dynamically reallocated according to current resource requirements and the allocation policy. Idle physical nodes can be switched off to minimize energy consumption. We propose an energy efficient resource allocation approach that utilises the capability of VMs live migration to reallocate resources dynamically based on application demand as well as when the workload is staic . The basic idea is to use a heuristic that is consolidating and rearranging the allocation in an energy efficient manner. Pure reactive resource allocation delays workload and

operates over time scale of a few minutes. Pure predictive resource allocation preserves long-term workload statistics besides envisaging and allocating for the next few hours. Therefore, hybrid resource allocation technique is used which combines predictive with reactive allocation techniques and accomplishes substantial improvements in: (a) meeting SLAs, (b) conserving energy, and (c) reducing provisioning costs and avoiding overutilization and underutilization of server.

II. VIRTUALIZATION TECHNOLOGY TO MINIMIZE ENERGY CONSUMPTION IN CLOUD DATA CENTERS

Virtualization technology allows one to create several VMs on a physical server thereby reducing the amount of hardware in use. Hardware virtualization hides the underlying computing system to present an abstract computing platform by using a hypervisor. In data centres, the number of physical machines can be reduced using virtualization by consolidating VMs onto shared servers which helps improve the efficiency of IT systems. The advantages are simple, it allows multiple virtual machines to be run on a single physical machine in order to provide more capability and increase the utilization level of the hardware. In-order to make the most out of Virtualization to save energy, the Cloud computing environment should comply with the following requirements:

- Virtualization of the infrastructure to support hardware and software heterogeneity, optimum resource utilization and simplify the resource provisioning.
- Application of VM migration to continuously adapt the allocation and quickly respond to changes in the workload.
- Ability to handle multiple applications with different SLA requirements owned by multiple users.
- Guaranteed meeting of the QoS requirements for each application.
- Support for different kind of applications, mixed workloads.
- Decentralization and high performance of the optimization algorithm to provide scalability and fault tolerance.
- Optimization considering multiple system resources, such as CPU, memory, disk storage and network interface.
- Intelligent consolidation and live migration mechanism improve the energy efficiency.

III.ENERGY EFFICIENCY

Energy efficiency refers to a reduction of energy used for a given service or level of activity, as defined by the World Energy Council. However, defining the energy efficiency of data center equipment is extremely difficult because it represents a complex system with a large number of components from various research areas such as computing, networking, management, and the like. Beloglazov et al. [13] define an energy model through static and dynamic power consumption, which deals only with energy waste while running idle. On the other hand, Avelar et al. define a difference between energy used by ICT and auxiliary equipment in order to measure energy losses by the latter. The issue of energy consumption across the ICT infrastructure such as data center is important and has received wide recognition in ICT sector. A new scalable design for efficient cloud computing infrastructure is required that can support the reduction in Green house Gas transmissions in general, and energy consumption in particular. The increase in ICT resource number and density has direct impact on the user expenditure as well as cooling and power management of data center infrastructure. In cloud computing, two popular schemes: (a) sleep scheduling and (b) resource virtualization had helped in improvement of energy efficiency and power consumption within the data center. The datacenter providers are now starting to realize the relationship between energy consumed by their ICT resources and GHG emissions. The three areas where energy is most consumed within a data center includes: (a) critical computational server providing CPU and storage functionalities; (b) cooling systems; and (c) power conversion units. There are two critical points where energy is not used in an efficient way but is instead lost or wasted. Both terms define inefficient use of energy from an agnostic point of view, where energy loss refers to energy brought to the system but not consumed for its main task (e.g., energy lost due to transport and conversion). This also includes energy used by supporting subsystems, such as cooling or lighting within a data center whose main task is the provision of cloud services. Energy waste refers to energy used by the system's main task but without useful output (e.g., energy used while running in idle mode). Both critical points are shown in Figure 1.

Therefore, two goals are defined for reducing energy loss and two goals for reducing energy waste, thus improving the energy efficiency:

L1. Energy not consumed by any subsystem— The first goal is minimizing a percentage of input energy that is not consumed by a subsystem. This can be done by implementing more efficient components (e.g., using more efficient power supply units for servers that leak less energy).

L2. Overhead of the supporting subsystems— the second goal is to reduce the overhead of supporting systems (i.e., systems that do not perform the main task of the system), for example, by implementing a single cooling unit for the entire cabinet instead of cooling each rack server separately.

W1. Idle run of the system— The third goal is to reduce idle run of the system and increase utilization or achieve zero energy consumption when no output is produced (i.e., during idle time). This also implies achieving a proportional increase of energy consumption with system output (e.g., to provide twice as much bandwidth, a network router requires twice the amount of energy or less).

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

W2. Redundant run of the system— the fourth goal is to minimize energy consumption where the system performs redundant operations. This can be done by implementing smart functions and subsystems, such as implementing optimized algorithm that does not require redundant steps to perform the same task.

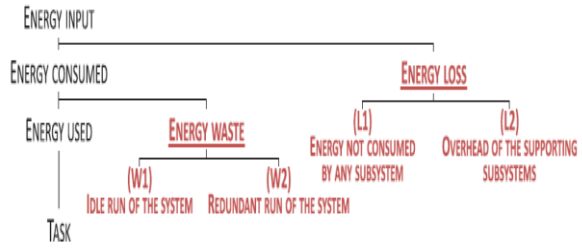

Figure1: Critical points within a system where energy is lost or wasted [14].

IV. CLOUD COMPUTING DATA CENTER DOMAINS

Cloud computing represents a novel and promising paradigm for managing and providing ICT resources to remote users. As the most cited definition of cloud computing, the U.S. National Institute of Standards and Technology (NIST) defines it as "a model that enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources, e.g., networks, servers, storage, applications, and services." It utilizes technologies such as virtualization, distributed computing, Service Oriented Architecture (SOA), and Service Level Agreements (SLAs) based on which different service types are offered. Regardless of its deployment or service model, cloud computing services are powered by large data centers comprised of numerous virtualized server instances and high-bandwidth networks, as well as of supporting systems such as cooling and power supplies. The listed equipment can be classified into two types, as shown in Figure 2; namely, hardware and software equipment [11].

Hardware includes both ICT equipment and supporting equipment within a data center, as defined in Avelar et al. ICT equipment includes Network and Server domains because they perform the main task of the data center and are the main focus of this survey. Domains such as Power supply, Cooling, and the Data center building itself are considered supporting equipment. Software equipment within a data center includes everything that runs on top of the ICT equipment. It includes two domains such as CMS and Appliances. Cloud Management Systems (CMS) are used to manage the entire data center and Appliances include software used by a user.

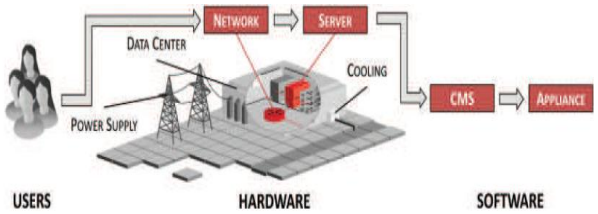

Figure 2: Cloud computing data center domains [14].

V. CLOUD COMPUTING INFRASTRUCTURE DOMAIN

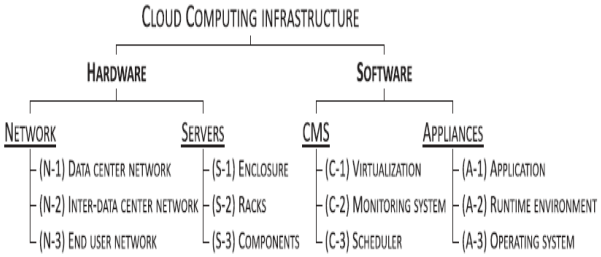

Figure3: Cloud computing infrastructure domains and related systems [14].

Cloud computing infrastructure domain can be classified into two types, as shown in Figure 3; namely, hardware and software equipment. Hardware includes both ICT equipment and supporting equipment within a data center which includes Network and Server domains because they perform the main task of the data center.

Software equipment within a data center includes everything that runs on top of the ICT equipment which consists of Cloud Management Systems (CMS) that are used to manage the entire data center and Appliances, which include software used by a user.

VI. NETWORK DOMAIN AND ITS ENERGY CASCADES

The network is a key enabling component for cloud computing since it allows communication between computing and storage resources and allows the end user to access them. The energy consumption of the Network domain consists of three main systems: the connections inside of a data center, the fixed network between data centers, and the end user network. Based on this breakdown, each system brings its own energy wastes and losses as given below:

—DCN Data center network (N-1): Within a data center, the energy consumption of the network currently accounts for up to 5% of its total energy consumption. As shown by Abts et al, network power accounts for approximately 20% of the total power when the servers are utilized at 100%. However, it goes up to 50% when utilization of servers decreases to 15%. As a result, this leads to an increasing share in energy consumption for these networks.

—D2D Inter-data center network (N-2): Connections between data centers are important for applications that run on a global scale, where instances that serve individual users are located in the data center closest to the end user but still need to communicate between each other.

—End user network (N-3): A connection to an end user who is accessing cloud services is usually made through a combination of wired and wireless networks. To reduce energy loss and waste, a number of actions can be taken.

These actions include:

—L1. Energy Loss1: Reducing the heat load of network equipment inside a data center (N-1) would reduce its

DOI 10.17148/IJARCCE.2016.5240

energy consumption and the consumption of its cooling subsystem as well.

—L2. Energy Loss2: Goal L1 also brings benefits to goal L2 by reducing heat load, a smaller cooling subsystem can be installed, which consumes less energy.

—W1. Energy Waste1: Today's network equipment is not energy proportional, and simply turning on a switch can consume over 80% of its max power. By implementing power saving modes, rate adaptation, or simply turning off unused ports, links, and switches inside a data center (N-1) would reduce idle energy consumption and therefore achieve this goal.

—W2. Energy Waste2: Achieving this goal depends mostly on how a network is used by the servers, as well as the Software and User domains. However, some optimization can still be done by observing communication patterns and reducing unnecessary traffic.

VII. SERVER DOMAIN AND ITS ENERGY CASCADES

The Server domain includes computing and storage servers as well as other components such as processors, memory, cabinets, and the like (but excluding communication equipment, which is part of the Network domain). As the second domain of IT equipment within a data center, its consumption contributes a large portion to the total energy consumption of a data center. Hence, improving server energy efficiency represents a top-priority task in the IT industry which is the burning research area both in industry and academics.

In a perfect data center, the Server domain, along with the Network domain, would consist only of hardware equipment that consumes energy. Therefore, an obvious goal of every data center owner is to reduce the consumption of all supporting hardware equipment because it represents an energy loss.

—Server Enclosure (S-1): Enclosures may differ depending on the type of cooling applied to a data center. The most common air-based cooling, based on Computer Room Air Conditioners (CRACs), requires enclosures to have air inlets and outlets on opposite sides. The second type of cooling is indirect liquid cooling.

—**Server Racks (S-2)**: The idle power consumption of a server can be more than 50% of its peak power consumption. Additionally, this includes a huge energy waste by running servers idle without any useful output or with low utilization in the 10–50% utilization range, which is usually the case in typical data centers [11].

—Server Components (S-3): The energy efficiency of server components drastically affects the overall efficiency of a server. Most focus on components that take a bigger slice of the total energy consumption, such as the CPU, which can consume more than a third of total server energy consumption. In addition to underutilized CPU cores that affect dynamic power consumption, caches can also be poorly used or underutilized, which adds to the static power consumption of a processor. Memory also creates energy overheads since it is built to provide high performance to meet increasing CPU demands and thus has grown in density, functionality, and scalability.

To mitigate all this energy loss and waste, a number of actions can be performed which include:

—L1.Energy Loss1: Reducing the heat load of server components such as the CPU fulfils this goal. This can be achieved by using more energy-efficient components and their architectures for CPU, memory, disk storage, etc.

—L2. Energy Loss2: Following goal L1, goal L2 provides additional energy savings by reducing energy consumed by supporting systems, such as cooling and power supplies inside the server enclosure (S-3) and the servers themselves (S-2). In addition to cooling and power supply systems, during idle run, subsystems such as cache can be turned off.

—W1. Energy Waste1: Using components that can automatically scale their power consumption based on current load would move toward achieving this goal; for example, using dynamic voltage and frequency scaling (DVFS)-capable CPUs that provide different P-states (power modes while being utilized), as well as sleep C-states (power modes while idle). Dharwar et al. [2012] provide an overview of these techniques along with power Capping[12]. The same applies memory and storage disks, which can be put into a low power state while idle.

—W2. Energy Waste2: As shown by many studies, bigger cache size does not necessarily mean a lower miss rate. Therefore, choosing the right size cache can decrease energy waste and achieve this goal.

VIII. RESOURCE ALLOCATION IN CLOUD DATA CENTER

Resource allocation is one of the challenges of cloud computing because end-users can access resources from anywhere and at any time. The resources in a cloud cannot be requested directly but can be accessed through SOAP/Restful web APIs that map requests for computations or storage aremapped to virtualized ICT resources (servers, blob storage, elastic IP, etc.). Since, cloud data center offer abundance of resources, the cloud computing model is able to support on-demand elastic resource allocation. In cloud computing paradigm, the key challenge is the allocation of resources among end-users having changing requests of resources based on their application usage patterns. The unpredictable and changing requests need to run on data center resources across the Internet. The aim of resource allocation for any particular cloud provider can be either optimize applications' QoS or improve resource utilization and energy efficiency. The main objective is to optimize QoS parameters (response time) that measures the efficiency of resource allocation regardless of the type of ICT resources allocated to end-users. Some of the challenges associated with energy efficient resource allocation policies are:

(a) Choosing workload type and interference between different workloads, such as resource usage, performance, and power consumption.

DOI 10.17148/IJARCCE.2016.5240

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

(b) Provisioning of resource allocation and utilization at run time by evaluating the possibility of centralized, federated, and standardized DataCenter resources.
(c) Improving asset utilization, network accessibility, power efficiency, and reduction in the time needed to recover from any failure.
(d) Improving cloud resources, topology, tools, and technologies by evaluating and fine tuning the cloud infrastructure layout.
(e) Increasing performance and the return on investment by assessing application inter-dependencies to facilitate resource consolidation.
(f) Supporting business security and flexibility for mission-critical applications through practical cloud infrastructure planning.

The definition of resource is very important as anything, such as CPU, memory, storage, bandwidth, and application can be termed an ICT resource in cloud computing landscape. The important characteristic of a resource unit is abstracted by the cost of operation and infrastructure.

The problem of resource allocation is quite complex and needs some assumptions including: (a) set of workflow tasks for resource requirements, (b) set of operational servers, (c) task consolidation meeting SLA, and (d) reduction in power wastage and resource usage costs. The resource allocation problem involves the appropriate provisioning and efficient utilization of available resources for applications to meet the QoS performance goals as per SLA. For cloud computing infrastructures, the service providers also need to track the changes in resource demands. Moreover, a cloud service provider allocates system resources to CPUs, and determines whether to accept incoming requests according to resource availability.

IX. TAXONOMY OF RESOURCE ALLOCATION TECHNIQUES IN CLOUD DATA CENTER

Resource allocation techniques with the focus on energy efficient resource management problem in data center clouds are given below. To do so, this section, makes a comparison based on the following dimensions: resource adaption policy, objective function, allocation method, and allocation operation.

A. RESOURCE ALLOCATION ADAPTATION POLICY

This dimension refers to the degree to which an energy-aware resource allocator is able to adapt to dynamic or uncertain conditions. The uncertainties arise from a number of factors including resource capacity demand (e.g., CPU, bandwidth, memory, and storage space), failures (e.g., failure of a network link and failure of the CPU hosting application instance), and user workload pattern (e.g., number of users and location). The resource adaption policy is classified into three categories: (a) predictive resource allocation adaption technique, (b) reactive technique, and (c) hybrid technique. Figure 4 depicts the pictorial representation of policy.

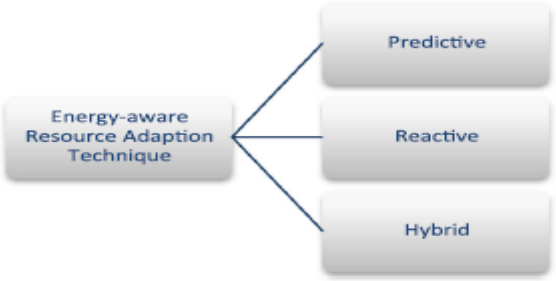

Figure 4: Resource adaption policy taxonomy

Monitoring activity of cloud resources involves dynamically profiling the QoS parameters related to hardware and software resources, the physical resources they share, and the applications running on them or data hosted on them. Monitoring services can help resource allocator as regards to: (i) keeping the cloud resources and application operating at peak energy efficiency levels; (ii) detecting the variations in the energy efficiency of resources and QoS delivered by hosted applications; and (iii) tracking the failure of resources and applications. The three resource adaption policies are given below:

(**a) Predictive Technique**: Using past knowledge-driven machine learning techniques, predictive resource allocation policy can dynamically anticipate and capture the relationship between applications QoS targets, energy efficiency objective function, and current hardware resource allocation and user workload patterns in order to adjust the resource allocation. The past knowledge is derived from the monitoring service, which continuously profiles information in a searchable database (e.g., MySQL and NoSQL databases). The resource capacity planning is done at prior and allocations are approximated based on resource performance models and application workload models. Workload prediction models forecast workload behaviour across applications in terms of CPU, storage, I/O, and network bandwidth requirements. On the other hand, resource performance model predict the performance of CPU, storage, and network resources based on their past performance history, and anticipated workload patterns. Predictive allocation handles the estimated base workload at coarse time scales (e.g., hours or days) maintaining long term workloads. The predictive resource allocation suffers from limitation when there is no sufficient workload and resource performance data available to train the machine learning technique. Predictable approaches can also fail under situations when the workload and resource performance data do not have any specific distributions. This affects the accuracy of prediction.

(b) Reactive Technique: Reactive techniques rely on monitoring the state of cloud resources and triggering hard-coded, pre-configured corrective actions when some specific even occurs such as utilization of the CPU resource reaches certain threshold or energy consumption of a CPU resource goes beyond threshold. The efficiency of reactive allocation depends on the ability to detect fluctuations. Reactive resource allocation adapts to service

DOI 10.17148/IJARCCE.2016.5240

requirements and resource availability, and optimize for long term resource allocation. Reactive policies react quickly to changes in workload demand but have limited significance because the policies suffer from issues such as: (a) lack of predictability, (b) instability, and (c) high provisioning costs. Pure reactive resource allocation delays workload and operates over time scale of a few minutes. Pure predictive resource allocation preserves long-term workload statistics besides envisaging and allocating for the next few hours.

(c) Hybrid Technique: Hybrid resource allocation combines predictive with reactive allocation techniques and accomplishes substantial improvements in: (a) meeting SLAs, (b) conserving energy, and (c) reducing provisioning costs. In hybrid resource allocation approach, a predictive (reactive) allocation switches the underline workload (handles any excess demand) at granular (finer) time scales. A coordinated management among predictive and reactive approaches achieves a significant improvement in energy efficiency without sacrificing performance . Hybrid allocation approaches outperforms predictive and reactive resource allocation strategies when performance, power consumption, and number of changes in resource allocation are considered.

B. OBJECTIVE FUNCTION

The objective function can be a mathematical expression, metric, or function that needs optimization with conditions subject to system constraints. Energy optimization objective function, which is also referred as cost function and energy function in would be either single or composite respecting to the number of parameters considered for optimization. For instance, a cost function which aims at minimizing energy consumption is considered as single objective function; hence if it deals with minimizing both energy consumption and SLA violation, the cost model would be composite. In cloud computing, for the increasing cost and shortage of power, an objective function is the measure of increase power usage for a resource allocation.

C. ENERGY EFFICIENT RESOURCE ALLOCATION IN CLOUD DATA CENTER

The energy conservation in hosting centers and server farms is of increasing importance. In the context of cloud computing, power-aware design techniques aim to maximize service-level performance measures (e.g. SLA violation rate, SLA accomplishment rate, waiting time, etc.) under power dissipation and power consumption constraints. Moreover, a power-aware technique can also help to reduce the energy cost. Power-aware strategies can be activated either in hardware or software level. For instance, dynamic component deactivation (DCD) strategy at hardware level is applied along with Advanced Configuration and Power Interface (ACPI) strategy at software level, since even with optimized hardware, poor software level design or optimization can lead to extensive power losses. Power-aware technologies either use low power energy-efficient hardware equipment (e.g., CPUs and power supplies) to reduce energy usage and peak power consumption, or reduce energy usage based on the knowledge of current resource utilization and application workloads. Power-aware scheduling process works at circuit, device, architectural, compiler, operating system, and networking layers. The most efficient and direct method is to use more power efficient component in the hardware design phase. Other approaches include developing algorithms for scaling down power or even turning down a system for unused resources.

D. RESOURCE ALLOCATION OPERATION

Resource allocation operation for optimizing energy efficiency of cloud resource can be classified into following categories:

(a) Service Migration: The transferring of process states and local data of an application component instance (e.g. web server and database server) to a new CPU resource (virtual machine container or virtual server) is called service migration. The service migration process enables dynamic load distribution by migrating processes from overloaded CPU or storage resources to less loaded ones, fault resilience by migrating processes from cloud resources that may have experienced a partial failure, eased system administration by migrating processes from the cloud resources that are about to be shut down or otherwise made unavailable, and data access locality by migrating processes closer to the source of some data. The major decision concerns of a service migration process are the time when a migration will occur, the selection process of the service which will migrate, and at which destination resource a service will move. Although, there are different power-aware algorithms for the host overload/ under load detection, CPU selection, and CPU placement, the service migration has still a lot of complexities. For instance, consider an application component migration from small instance to medium instance or large instance. These instances vary in their hardware configuration such as RAM, cores, local storage size, addressing bits and I/O performance.

(b) Service Shutdown: Service shutdown refers to automatic switching/powering off the system, hardware component, or network on the failure or abnormal termination of the previously active resource allocation. Service shutdown can be automatic or might require human intervention. One of the major reasons for service shutdown is to conserve energy. However, before shutting down a server, all the running services have to be consolidated and migrated to other nodes power and migration cost aware application placement algorithms in virtualized systems.

X. RELATED WORK

Currently, resource allocation mechanisms which are widely used in data centres include load balancing, round robin and greedy algorithms. The load balancing module tries to distribute the workload evenly over the computing nodes of the system. In the round robin algorithm, the servers are in a circular list. There is a pointer pointing to the last server that received the previous request. The

system sends a new resource allocation to the next node with enough physical resources to handle the request. Having the same set up as the round robin algorithm, the greedy algorithm continues to send new resource allocation request to the same node until it is no longer has enough physical resources, then the system goes to the next one. The work in [2] studies general energy saving approaches. Saving energy in ICT divides into two main fields, saving energy for network [3,4] and saving energy for computing nodes [5,6]. Related to saving energy in cloud data centre, the work in [7] adjusts the working mode of the server according to load to save energy. In [8], the authors focus on saving energy for PaaS (Platform as a Service) cloud data centre. Our work is different from previous in two main ways. We focus on IaaS scenario and use moving workload as the main method. [9] is the most closely work with our work. However, in [9] the authors use the predefined MIPS to present the capacity and the load which is quite difficult to the user. In related study, a scheme for selecting energy efficient allocation of virtual machines in cloud data center. This scheme considers the maximum and minimum utilization threshold value. If the utilization of CPU for a host falls below the lower threshold, all VMs have to be migrated from this host and the host has to be switched off in order to eliminate idle power consumption. If the utilization goes over the upper threshold, some VMs have to be migrated from the host to reduce utilization to prevent potential SLA violation.

We propose three policies for choosing VMs that have to be migrated from the host: (1) Minimization of Migrations(MM) – migrating the least number of VMs to minimize migration overhead; (2) Highest Potential Growth (HPG) –migrating VMs that have the lowest usage of CPU relatively to requested in order to minimize total potential increase of the utilization and SLA violation; (3) Random Choice (RC)– migrating the necessary number of VMs by picking them according to a uniformly distributed random variable[10]. Our work follows the similar line of fashion as given in[10] to achieve energy efficient resource allocation in cloud data centers while providing QoS and meeting desired SLA.

XI. PROBLEM DEFINITION

In related study, a scheme for selecting energy efficient allocation of virtual machines in cloud data center. Proposed scheme consider the maximum and minimum utilization threshold value. If the utilization of CPU for a host falls below the minimum threshold, all VMs have to be migrated from this host and the host has to be switched off in order to eliminate idle power consumption. If the utilization goes over the maximum threshold, some VMs have to be migrated from the host to reduce utilization to prevent potential Service Level Agreements violation. Further for migrating, it uses minimization of migrations to reduce migration overhead. This paper presents a hybrid energy efficient resource allocation technique which combines predictive with reactive allocation techniques and accomplishes substantial improvements in: (a) meeting SLAs, (b) conserving energy, and (c) meeting static and dynamic resource allocation. In this paper we propose energy-aware allocation heuristics provision data center resources to client applications in a way that utilises the capability of VMs live migration to reallocate resources dynamically and improves energy efficiency of the data center, while delivering the negotiated Quality of Service (QoS). The basic idea is to use a heuristic that is consolidating and rearranging the allocation of resources in an energy efficient manner. Load balancing process engine will judge the threshold variation and if the value goes down or show variation in upper and lower bound of the decided threshold for resource allocation. Almost three iterations will be done for discussed scheme and values obtained from these cycled processes are stored and algorithm will judge a average value for migration. We will run the each cycle with variation in intial resources allocation values which will provide variation in number of migration required for particular task completion. This process will be reapeated and results with number of migration required in different scenarios will be fetched. Value of optimized migration is obtained from best of three scenarios and after choosing this scheme, final migration resources will be assigned. Upper and lower bound limits will be decided as tolerant values in general like 90% higher and 15% lower bound values. Increase and decrease to specific decided values will migrate or shut down the virtual machines. This process will be helpful to minimize total potential increase of the CPU utilization, RAM and SLA violation. For validation of our proposed work, we will use CloudSim.

XII. RESEARCH METHODOLOGY

This paper presents a hybrid energy efficient resource allocation technique which combines predictive with reactive allocation techniques and accomplishes substantial improvements in: (a) meeting SLAs, (b) conserving energy, and (c) meeting static and dynamic resource allocation. In this paper we propose energy-aware allocation heuristics provision data center resources to client applications in a way that utilises the capability of VMs live migration to reallocate resources dynamically and improves energy efficiency of the data center, while delivering the negotiated Quality of Service (QoS). Our research will start with study of parallel computing management in virtual cloud environment based on cloud computing for virtualization in following steps:

1st Phase: This will be the initial stage for the whole process and will contain the basic functionality and collection of information (virtual simulation, basic virtualization functions etc). Layout for comparison will be done in this phase.

2nd Phase: In this stage we will implement the basic scenario for parallel processing structure based on resources allocation scheme.

3rd Phase: In this stage we will create the migration instructions and checking of the process based on prediction algorithm to find the particular resources which need to be migrated.

4th Phase: This the prime stage of the experimentation in

which we will apply the prediction based algorithm which provides us the details of resources requirement for various users. The resource utilization of the complete process will be judged based on the threshold value of the utilization of the multiprocessor environment in parallel communication. The basic function of the selection of the threshold will be done by the selection engine which will be deciding the threshold according to the requirement of the system and can be assign the threshold values dynamically. Load balancing process engine will judge the threshold variation and if the value goes down or show variation in upper and lower bound of the decided threshold for resource allocation. Almost three iterations will be done for discussed scheme and values obtained from these cycled processes are stored and algorithm will judge a average value for migration. We will run the each cycle with variation in initial resources allocation values which will provide variation in number of migration required for particular task completion. This process will be repeated and results with number of migration required in different scenarios will be fetched. Value of optimized migration is obtained from best of three scenarios and after choosing this scheme, final migration resources will be assigned. Upper and lower bound limits will be decided as tolerant values in general like 90% higher and 15% lower bound values. Increase and decrease to specific decided values will migrate or shut down the virtual machines. The host needs to switch off in case of migration to save the power consumption of the server which in turns provide less heating effect. In case the value crosses the upper bound rapidly then engine need to migrate the partial processes to other host server (which is already judge for resource utilization by the prediction algorithm in the first place). Further for better execution time, migration of the virtual machine will be on lesser side which in turn could also decrease the overall overhead. This research provides the similar line of process as described in related study [1].

5th Phase: Final stage will be for the comparison of the proposed work with already existing work.

XIII. CONCLUSION AND FUTURE WORK

In recent years, energy efficient resource allocation of data center resources has evolved as one the critical research issue. We have identified power and energy inefficiencies in hardware and software, and categorized existing techniques in this paper. We have presented the design, implementation, and evaluation of a resource management system for cloud computing services. Our system multiplexes virtual to physical resources adaptively based on the changing demand.

We combine VMs with different resource characteristics appropriately so that the capacities of servers are well utilized. Our methodology achieves both overload avoidance and energy efficient computing for systems with multi-resource constraints. This paper presents a hybrid energy efficient resource allocation technique which combines predictive with reactive allocation techniques and accomplishes substantial improvements in: (a) meeting SLAs, (b) conserving energy, and (c) meeting static and dynamic resource allocation. In this paper we propose energy-aware allocation heuristics provision data center resources to client applications in a way that utilises the capability of VMs live migration to reallocate resources dynamically and improves energy efficiency of the data center, while delivering the negotiated Quality of Service (QoS). The basic idea is to use a heuristic that is consolidating and rearranging the allocation of resources in an energy efficient manner. For the future research work we suggest to develop intelligent techniques to manage the network resources efficiently.

REFERENCES

[1] Zhen Xiao, ‘‘Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment’’, IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 6, JUNE 2013.

[2] A. Berl, E. Gelenbe, M. di Girolamo, G. Giuliani, H. de Meer, M.-Q. Dang, and K. Pentikousis. Energy-Efficient Cloud Computing. The Computer Journal, 53(7), September 2010, doi:10.1093/comjnl/bxp080.

[3] E. Gelenbe and C. Morfopoulou, A Framework for Energy Aware Routing in Packet Networks. accepted for publication in The Computer Journal.

[4] E. Gelenbe and T. Mahmoodi, Energy-Aware Routing in the Cognitive Packet Network In International Conference on Smart Grids, Proceeding of Energy 2011 conference, 2011.

[5] T. Heath, B. Diniz, E. V. Carrera, W. M. Jr., and R. Bianchini. Energy conservation in heterogeneous server clusters. Proc. of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming - PPoPP'05, 2005, pp. pages 186-195.

[6] C. Xian , Y. H. Lu, Dynamic voltage scaling for multitasking real-time systems with uncertain execution time, Proc. of the 16th ACM Great Lakes symposium on VLSI, 2006.

[7] T. V. Do ,Comparison of Allocation Schemes for Virtual Machines in Energy-Aware Server Farms, The Computer Journal, 2011.

[8] J. Leverich and C. Kozyrakis, On the Energy (In)efciency of Hadoop Clusters, SIGOPS, 2010.

[9] A.Beloglazov and R. Buyya, Energy Efficient Resource Management in Virtualized Cloud Data Centres, Proceeding of the 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010.

[10] A. Beloglazov and R. Buyya,” Energy Efficient Allocation of Virtual Machines in Cloud Data Centers”, 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing,2010.

[11] Urs Hoelzle and Luiz Andre Barroso. 2013. The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines (2nd ed.). Morgan and Claypool.

[12] D. Dharwar, S. S. Bhat, V. Srinivasan, D. Sarma, and P. K. Banerjee. 2012. Approaches towards energy- efficiency in the cloud for emerging markets. In Proceedings of the 2012 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM’12). 1–6. DOI:http://dx.doi.org/10.1109/ CCEM.2012.6354599.

[13] Anton Beloglazov, Rajkumar Buyya, Young Choon Lee, and Albert Y. Zomaya. 2011. A taxonomy and survey of energy-efficient data centers and cloud computing systems. Advances in Computers 82 (2011), 47–111.

[14] Toni Mastelic, Ariel Oleksiak, Holger Claussen, Ivona Brandic, Jean-Marc Pierson, et al.. Cloud computing: survey on energy efficiency. ACM Computing Surveys, Association for Computing Machinery, 2015, Vol. 47 (n 2), pp. 1-36.

.