Energy-Efficient Cloud Computing: A Vision, Introduction, and Open Challenges

[1]Ms.Jayshri Damodar Pagare, [2]Dr.Nitin A Koli

[1] Computer science and Engineering Department, Jawaharlal Nehru Engineering College,
Dr.Babasaheb Ambedkar Marathwada University
Aurangabad, Maharashtra 431 003, India

[2] Head Computer Center , Sant Gadge Baba Amravati University
Amravati, Maharashtra 444 602, India

Abstract

Cloud computing is offering utility-oriented IT services to users worldwide. Based on a pay-as-you-go model, it enables hosting of pervasive applications from consumer, scientific, and business domains. In recent years, energy efficiency has emerged as one of the most important design requirements for modern computing systems, such as data centers and Clouds, as they continue to consume enormous amounts of electrical power , contributing to high operational costs and carbon footprints to the environment .The virtualization technology has advanced the area by introduction of a very effective power saving technique.
This paper presents vision, introduction to cloud computing and challenges for energy-efficient management of Cloud computing environments. We focus on the development of energy efficient algorithms that work to boost data center energy efficiency and performance.This paper proposes development of Virtualization technology that provides some unique opportunity for better resource utilization and develop a software platform that supports the energy efficient management and allocation of Cloud data center resources.

Keywords: *Iaas, SaaS, PaaS, Virtualization.*

1. Introduction

Cloud Computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the datacenters that provide those services. The services themselves have long been referred to as Software as a Service (SaaS). The datacenter hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is Utility Computing. We use the term Private Cloud to refer to internal datacenters of a business or other organization, not made available to the general public. Thus, Cloud Computing is the sum of SaaS and Utility Computing, but does not include Private Clouds. People can be users or providers of SaaS, or users or providers of Utility Computing. From a hardware point of view, three aspects are new in Cloud Computing.

1. The illusion of infinite computing resources available on demand, thereby eliminating the need for Cloud Computing users to plan far ahead for provisioning.
2. The elimination of an up-front commitment by Cloud users, thereby allowing companies to start small and increase hardware resources only when there is an increase in their needs.
3. The ability to pay for use of computing resources on a short-term basis as needed (e.g., processors by the hour and storage by the day) and release them as needed, thereby rewarding conservation by letting machines and storage go when they are no longer useful.

The Cloud computing model leverages virtualization of computing resources allowing customers to provision resources on-demand on a pay-as-you-go basis [1]

A. Significance

The main reason why anyone should use cloud computing paradigm is scalability, particularly for research projects that require vast amounts of storage or processing capacity for a limited time. Cloud computing presents IT organizations with a fundamentally different model of operation, one that takes advantage of maturity of web applications and networks and the rising interoperability of computing system to provide IT services. Cloud providers specialize in particular applications and services, and this expertise allows them to efficiently manage upgrades and maintenance , backups, disaster recovery, and failover function. As a result, consumers of cloud services may see increased reliability, even as costs decline due to economics of scale and other production factors. With cloud computing, organizations can monitor current needs and make on-the-fly adjustments to increase or decrease capacity, accommodating spikes in demand without paying for unused capacity during slower times.

Aside from the potential to lower costs, colleges and universities gain the flexibility of being able to respond quickly to requests for new services by purchasing them from the cloud. Cloud computing encourages IT organizations and providers to increase standardization of protocols and processes so that the many pieces of the cloud computing model can interoperate properly and efficiently. Some companies have built data centers near resources of renewable energy, such as wind farms and hydroelectricity facilities , and cloud computing affords access to these providers of " green IT". Cloud computing allows college and university IT providers to make IT costs transparent and thus much consumption of IT services to those who pay for such services.

B. Advantages

Cloud computing provides numerous economic advantages

For clients:

- No upfront commitment in buying/leasing hardware
- Can scale usage according to demand
- Barriers to entry lowered for startups

For providers:

- Increased utilization of datacenter resources

C. Advantages of virtualization

- Isolation
- consolidation
- Migration

D. Limitations

- Increased attack surface

Entity outside the organization now stores and computes data, and so Attackers can now target the communication link between cloud provider and client

- Auditability and forensics

Difficult to audit data held outside organization in a cloud.

Forensics also made difficult since now clients don't maintain data locally

- Security

E. Application

Cloud computing can play a significant role in a variety of areas including internal pilots, innovations, virtual worlds, e-business, social networks, and search. The main consumer of cloud computing are small companies and startups that don't have a legacy of IT investments to manage.

2. Objectives

The main objective of this work is to present our vision, discuss open research challenges in energy-aware resource management, and develop efficient policies and algorithms for virtualized data centers so that Cloud computing can be a more sustainable and eco-friendly mainstream technology to drive commercial, scientific, and technological advancements for future generations. Specifically, our work aims to:

- Develop algorithms for energy-efficiency.
- Explore open research challenges in energy-efficient resource management for virtualized Cloud data centers.

A. Goals

- Learn about latest research
- Adapt well known techniques for virtualization
- Perform new research

B. Expected Innovation

Design and implementation of simulators and testbeds for evaluation of mentioned algorithm

C. Scope

Virtualization of computer resources is widely recognized as an important revolution in computing industry over the last decade. Virtualization technology provides some unique opportunity for better resource utilization and more effective server and application consolidation.

3. Motivation

Cloud Computing, the long-held dream of computing as a utility, has the potential to transform a large part of the IT industry, making software even more attractive as a service and shaping the way IT hardware is designed and purchased. Developers with innovative ideas for new Internet services no longer require the large capital outlays in hardware to deploy their service or the human expense to operate it. They need not be concerned about over provisioning for a service whose popularity does not meet their predictions, thus wasting costly resources, or under provisioning for one that becomes wildly popular, thus missing potential customers and revenue. Moreover, companies with large batch-oriented tasks can get results as quickly as their programs can scale, since using 1000 servers for one hour costs no more than using one server for 1000 hours. This elasticity of resources, without paying a premium for large scale, is unprecedented in the history of IT[2].

The rapid growth in demand for computational power driven by modern service applications combined with the shift to the Cloud computing model have led to the establishment of large-scale virtualized datacenters. Such data centers consume enormous amounts of electrical energy resulting in high operating costs and carbon dioxide emissions. Dynamic consolidation of virtual machines (VMs) using live migration and switching idle nodes to the sleep mode allow Cloud providers to optimize resource usage and reduce energy consumption. However, the obligation of providing high quality of service to customers leads to the necessity in dealing with the energy-performance trade-off, as aggressive consolidation may lead to performance degradation. Due to the variability of workloads experienced by modern applications, the VM placement should be optimized continuously in an online manner.

4. Details of the proposed implementation

The research work is planned to be followed by the development of a software platform that supports the energy efficient management and allocation of Cloud data center resources. we will extensively reuse existing Cloud middleware and associated technologies. We will leverage third party Cloud technologies and services offerings including (a) VM technologies, such as open-source Xen and KVM, and commercial products from VMware; (b) Amazon's Elastic Compute Cloud (EC2), Simple Storage Service (S3), and Microsoft's Azure.

As the target system is a generic Cloud computing environment, it is essential to evaluate it on a large-scale virtualized data center infrastructure. However, it is difficult to conduct large-scale experiments on a real infrastructure, especially when it is necessary to reproduce the experiment with the same conditions to compare different algorithms. Therefore, a simulation has been chosen as a way to evaluate the proposed algorithms. The CloudSim toolkit 2.0 [3] has been chosen as a simulation platform, as it is a modern simulation framework aimed at Cloud computing environments. In contrast to alternative simulation toolkits (e.g. SimGrid, GangSim), it allows the modeling of virtualized environments, supporting on-demand resource provisioning, and their management. It has been extended in order to enable power-aware simulations and dynamic workloads, as the core framework does not provide these capabilities. The implemented extensions have been included in the 2.0 version of the CloudSim toolkit.

Recently, Yahoo and HP have led the establishment of a global Cloud computing testbed, called Open Cirrus, supporting a federation of data centers located in 10 organizations [4]. Building such experimental environments is expensive and hard to conduct repeatable experiments as resource conditions vary from time to time due to its shared nature. Also, their accessibility is limited to members of this collaboration. Hence, simulation environments play an important role.

The security and integrity of the VMs in the cloud becomes a major concern as more organizations turn to cloud virtualization solutions [5]. We wish to place a level of trust in the VM that it will perform its intended task without compromise or lose the data it is processing.

5. Review of literature

What really is **Cloud Computing**

Cloud computing is a new computing paradigm, involving data and/or computation outsourcing, with

- Infinite and elastic resource scalability
- On demand "just-in-time" provisioning
- No upfront cost … pay-as-you-go

That is, use as much or as less you need, use only when you want, and pay only what you use

Cloud

A cloud is a pool of virtualized computer resources. A cloud can:

- Host a variety of different workloads, including batch-style back-end jobs and interactive, user-facing applications
- Allow workloads to be deployed and scaled-out quickly through the rapid provisioning of virtual machines or physical machines
- Support redundant, self-recovering, highly scalable programming models that allow workloads to recover from many unavoidable hardware/software failures
- Monitor resource use in real time to enable rebalancing of allocations when needed

Cloud computing means **selling "X as aservice"**

IaaS: Infrastructure as a Service

- Selling virtualized hardware

PaaS: Platform as a service

- Access to a configurable platform/API

SaaS: Software as a service

- Software that runs on top of a cloud

Cloud computing is a pay-per-use model for enabling available, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, services) that can

be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is comprised of five **key characteristics,** three **delivery models**, and four **deployment models**.

Virtualization
Virtualization is the process of decoupling hardware from the operating system on a physical machine. A *Virtual Machine (VM)* is the virtualized representation of a physical machine that is run and maintained on a host by a software virtual machine monitor or *hypervisor.*

A. Cloud Computing Architecture

NIST (National Institute of Standards and Technology) is a well accepted institution all over the world for their work in the field of Information Technology. I shall present the working definition provided by NIST of Cloud Computing. NIST defines the Cloud Computing architecture by describing five essential characteristics, three cloud services models and four cloud deployment models (Cloud Security Alliance, 2009, p14).

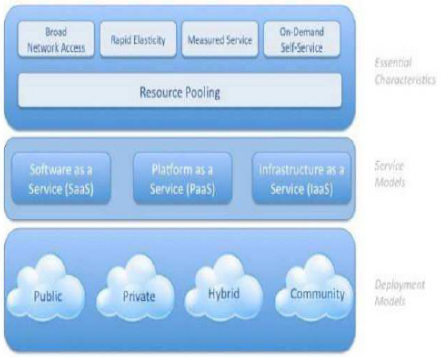

*Figure 1 - Visual model of NIST Working Definition of Cloud Computing (*Cloud Security Alliance, 2009, p14*)*

5.1 Key Characteristics

a) On-demand self-service
A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed without requiring human interaction with each service's provider.

b) Ubiquitous network access
Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).

c) Location independent resource pooling
The provider's computing resources are pooled to serve all consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. The customer generally has no control or knowledge over the exact location of the provided resources. Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.

d) Rapid elasticity
Capabilities can be rapidly and elastically provisioned to quickly scale up and rapidly released to quickly scale down. To the consumer, the capabilities available for rent often appear to be infinite and can be purchased in any quantity at any time.

e) Pay per use
Capabilities are charged using a metered, fee-for-service, or advertising based billing model to promote optimization of resource use. Examples are measuring the storage, bandwidth, and computing resources consumed and charging for the number of active user accounts per month. Clouds within an organization accrue cost between business units and may or may not use actual currency.

Note: Cloud software takes full advantage of the cloud paradigm by being service oriented with a focus on statelessness, low coupling, modularity, and semantic interoperability.

5.2 Delivery Models

f) Cloud Software as a Service(SaaS)
The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure and accessible from various client devices through a thin client interface such as a Web browser (e.g., web-based email). The consumer does not manage or control the underlying cloud infrastructure, network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

g) Cloud Platform as a Service(PaaS)
The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created applications using programming languages and tools supported by the provider (e.g., java, python, .Net). The consumer does not manage or control the underlying cloud infrastructure,

network, servers, operating systems, or storage, but the consumer has control over the deployed applications and possibly application hosting environment configurations.

h) Cloud Infrastructure as a Service(IaaS)

The capability provided to the consumer is to rent processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly select networking components (e.g., firewalls, load balancers).

5.3 Deployment Models

i) Private cloud

The cloud infrastructure is owned or leased by a single organization and is operated solely for that organization.

j) Community cloud

The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations).

k) Public cloud

The cloud infrastructure is owned by an organization selling cloud services to the general public or to a large industry group.

l) Hybrid cloud

The cloud infrastructure is a composition of two or more clouds (internal, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting).

Each deployment model instance has one of two types: internal or external. Internal clouds reside within an organizations network security perimeter and external clouds reside outside the same perimeter.

Note 1: Cloud computing is still an evolving paradigm. Its definitions, use cases, underlying technologies, issues, risks, and benefits will be refined in a spirited debate by the public and private sectors. These definitions, attributes, and characteristics will evolve and change over time.

Note 2: The cloud computing industry represents a large ecosystem of many models, vendors, and market niches. This definition attempts to encompass all of the various cloud approaches.

6. Challenges

Virtualization addresses many of the challenges faced by enterprises in using Cloud based services including:

- How to easily migrate applications and workloads to the Cloud ?
- How do you ensure business policies (SLA, security, backup etc.) move with the application to the Cloud ?
- How do you federate your workload between your own datacenter and external cloud based resources ?
- How do you rapidly provision new applications in an automated manner ?
- How can you monitor and manage resources scattered across multiple locations ?
- How do you allocate resources for maximum resource utilization?
- How do you lower operating costs by minimizing power and cooling cost?

7. Related work

One of the first works, in which power management has been applied in the context of virtualized data centers, has been done by Nathuji and Schwan [6]. The authors have proposed an architecture of a data center's resource management system where resource management is divided into local and global policies. At the local level the system leverages the guest OS's power management strategies. The global manager gets the information on the current resource allocation from the local managers and applies its policy to decide whether the VM placement needs to be adapted. However, the authors have not proposed a specific policy for automatic resource management at the global level. They have investigated the problem of power-efficient resource management in large-scale virtualized data centers. This is the first time when power management techniques have been explored in the context of virtualized systems. The authors have pointed out the following benefits of virtualization: improved fault and performance isolation between applications sharing the same resource; ability to relatively easy move VMs between physical hosts applying live or offline migration; and support for hardware and software heterogeneity.

Cardosa et al. [7] have proposed an approach for the problem of power-efficient allocation of VMs in virtualized heterogeneous computing environments. They have leveraged the min, max and shares parameters of Xen's VMM, which represent minimum, maximum and proportion of the CPU allocated to VMs sharing the same resource. However, the approach suits only enterprise environments as it does not support strict SLAs and requires the knowledge of application priorities to define the shares parameter. Other limitations are that the allocation of VMs is not adapted at run-time (the allocation is static). They have presented several techniques for addressing the sharing-aware VM allocation problem. Hypervisor distributes resources among VMs according to a sharing-based mechanism, when the minimum and maximum amount of resources that can be allocated to a VM are specified.

Jung et al. [8], [9] have investigated the problem of dynamic consolidation of VMs running a multi-tier web-application using live migration, while meeting SLA requirements. The SLA requirements are modeled as the response time precomputed for each type of transactions specific to the web-application. A new VM placement is produced using bin packing and gradient search techniques. The migration controller decides whether there is a reconfiguration that is effective according to the utility function that accounts for the SLA fulfillment. However, this approach can be applied only to a single web-application setup and, therefore, cannot be utilized for a multitenant IaaS environment.

Verma et al. [10] have formulated the problem of power-aware dynamic placement of applications in virtualized heterogeneous systems as continuous optimization: at each time frame the placement of VMs is optimized to minimize power consumption and maximize performance.

Kusic et al. [11] have investigated the problem of minimizing both power consumption and SLA violations for online services in virtualized data centers using a limited look-ahead control.[12]They have explored the problem of power- and performance-efficient resource management in virtualized computing systems. The problem is narrowed to the dynamic provisioning of VMs for multitiered web applications according to the current workload (number of incoming requests). The SLA for each application are defined as the request processing rate. The clients pay for the provided service and receive a refund in a case of violated SLA as a penalty to the resource provider. The objective is to maximize the resource provider's profit by minimizing both power consumption and SLA violation. The problem is stated as a sequential optimization and addressed using the limited lookahead control (LLC).

In addition, many studies have focused on power-aware real-time applications in clusters. Rusu et al. [13] have developed a QoS-aware power management scheme by combining cluster-wide (On/ Off) and local (DVFS) power management techniques in the context of heterogeneous clusters. The front-end manager decides which servers should be turned on or off for a given system load, while the local manager reduces power consumption using DVFS.

Furthermore, recent work on implementing real-time VMs [14, 15] assures real-time services (e.g. real-time CPU allocation, real-time I/O) of a VM.

Song et al. [16] have proposed resource allocation to applications according to their priorities in multi application virtualized clusters. They have studied the problem of the efficient resource allocation in multiapplication virtualized data centers. The objective is to improve the utilization of resources leading to the reduced energy consumption. To ensure the QoS, the resources are allocated to applications proportionally according to the application priorities. Each application can be deployed using several VMs instantiated on different physical nodes.

Meisner et al. [17] have proposed an approach for power conservation in server systems based on fast transitions between active and low-power states. The goal is to minimize power consumption by a server while it is in an idle state.

Stillwell et al. [18] have studied the problem of the resource allocation for HPC applications in virtualized homogeneous clusters. The objective is to maximize the resource utilization, while optimizing user-centric metric that encompasses both performance and fairness, which is referred to as the yield. The idea is to design a scheduler focusing on a user-centric metric. The yield of a job is "a fraction of its maximum achievable compute rate that is achieved." A yield of 1 means that the job consumes computational resources at its peak rate.

8. Conclusion

This paper has reviewed the virtualization in cloud computing for energy saving. We have surveyed the contributions that are available in this area from recent

research. We propose that cloud computing with virtualization as a way to achieve the main sources of energy consumption, and the significant trade-offs between performance, QoS and energy efficiency.

References

[1] R. Buyya et al. Market-oriented cloud computing: Vision, hype,and reality for delivering it services as computing utilities. In Proc. of the 10th IEEE Intl. Conf. on High Performance Computing and Communications (HPCC'08), 2008.

[2] Armbrustet al., Above the Clouds: A Berkeley View of Cloud Computing, UC Berkeley Tech Report UCB/EECS-2009-28, February 2009.

[3] R. N. Calheiros et al. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and Experience, Wiley Press, NY, USA, 2010.

[4] A. Avetisyan et. al. Open Cirrus: A Global Cloud Computing Testbed. IEEE Computer, April 2010.

[5] S. Berger, R. Cáceres, D. Pendarakis, et al., "TVDc: Managing Security in the Trusted Virtual Datacenter," ACM SIGOPS Operating Systems Review, vol. 42,no. 1, pp. 40-47, January, 2008.

[6] Nathuji R, Schwan K. Virtual power: Coordinated power management in virtualized enterprise systems. ACMSIGOPS Operating Systems Review 2007; 41(6):265–278.

[7] Cardosa M, Korupolu M, Singh A. Shares and utilities based power consolidation in virtualized server environments. Proceedings of the 11th IFIP/IEEE Integrated Network Management (IM 2009), Long Island, NY,USA, 2009

[8] Jung G, Joshi KR, Hiltunen MA, Schlichting RD, Pu C. Generating adaptation policies for multi-tier applications in consolidated server environments. Proceedings of the 5th IEEE International Conference on Autonomic Computing(ICAC 2008), Chicago, IL, USA, 2008; 23–32.

[9] Jung G, Joshi KR, Hiltunen MA, Schlichting RD, Pu C. A cost-sensitive adaptation engine for server consolidation of multitier applications. Proceedings of the 10th ACM/IFIP/USENIX International Conference on Middleware (Middleware 2009), Urbana Champaign, IL, USA, 2009; 1–20

[10] A. Verma, P. Ahuja, A. Neogi, pMapper: Power and migration cost aware application placement in virtualized systems,in: Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware, Springer, 2008, pp. 243{264.

[11] Kusic D, Kephart JO, Hanson JE, Kandasamy N, Jiang G.Power and performance management of virtualized computing environments via lookahead control. In Proc. of 5th IEEE Intl. Conf. on Autonomic Computing (ICAC 2008), pages 3–12. Chicago, USA, June 2008.

[12] D. Kusic, J.O. Kephart, J.E. Hanson, N. Kandasamy, G. Jiang, Power and performance management of virtualized computing environments via lookahead control, Cluster Comput. 12 (1) (2009) 1–15.

[13] Rusu C, Ferreira A, Scordino C, Watson A, Melhem R, MossenD. Energy-efficient real-time heterogeneous server clusters. In Proc. of the 12th IEEE Real-Time and Embedded Technology and Applications Symposium, pages 418–428. San Jose, USA, April 2006.

[14] Virtual Logicx Real-Time Virtulalization and VLX. VirtualLogix, http://www.osware.com.

[15] Yoo S, Park M, Yoo C. A step to support real-time in a virtual machine monitor. In Proc. of 6th IEEE Consumer Communications and Networking Conference. Las Vegas, USA, January 2009.

[16] Y. Song, H. Wang, Y. Li, B. Feng, Y. Sun, Multi-Tiered On-Demand resource scheduling for VM-Based data center, in: Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2009), Shanghai, China, 2009, pp. 148–155.

[17] D. Meisner, B.T. Gold, T.F. Wenisch, PowerNap: eliminating server idle power, ACM SIGPLAN Notices 44 (3) (2009) 205–216.

[18] M. Stillwell, D. Schanzenbach, F. Vivien, H. Casanova, Resource allocation using virtual clusters,in: Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid(CCGrid 2009), Shanghai, China, 2009, pp. 260–267.

First Author Jayshri Pagare is currently working as Assistant Professor in Department of Computer Science and Engineering, Jawaharlal Nehru Engineering College, Aurangabad.

Second Author Dr.Nitin A Koli is currently working as Head, Computer Center ,at Amravati University.