ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

# Load Balancing Policies for the Cloud Ecosystem Which Are Energy Responsive

**M.Parimala**
**Associate Professor**
**Department of CSE**
**Tirumala Engineering College.**

**J.Navaneetha**
**Associate Professor**
**Department of CSE**
**Tirumala Engineering College.**

**Muthe Swapna**
**M.Tech Student**
**Department of CSE**
**Tirumala Engineering College.**

*ABSTRACT*

***In this paper we introduce an energy-aware operation model used for load balancing and application scaling on a cloud. The basic philosophy of our approach is defining an energy-optimal operation regime and attempting to maximize the number of servers operating in this regime. Idle and lightly-loaded servers are switched to one of the sleep states to save energy. The load balancing and scaling algorithms also exploit some of the most desirable features of server consolidation mechanisms discussed in the literature.***

## INTRODUCTION

### What is cloud computing?

**Cloud computing** is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services. These services typically provide access to advanced software applications and high-end networks of server computers.


### How Cloud Computing Works?

The goal of cloud computing is to apply traditional supercomputing, or high-performance computing power, normally used by military and research facilities, to perform tens of trillions of computations per second, in consumer-oriented applications such as financial portfolios, to deliver personalized information, to provide data storage or to power large, immersive computer games.

The cloud computing uses networks of large groups of servers typically running low-cost consumer PC technology with specialized connections to spread data-processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together. Often, virtualization techniques are used to maximize the power of cloud computing.

### Characteristics and Services Models:

The salient characteristics of cloud computing based on the definitions provided by the National Institute of Standards and Terminology (NIST) are outlined below:

### On-demand self-service:

A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.

### Broad 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).

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

**Resource pooling:**
The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location-independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or data center). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.

**Rapid elasticity:**
Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

**Measured service:**
Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be managed, controlled, and reported providing transparency for both the provider and consumer of the utilized service.

## EXISTING SYSTEM:

An important strategy for energy reduction is concentrating the load on a subset of servers and, whenever possible, switching the rest of them to a state with low energy consumption. This observation implies that the traditional concept of load balancing in a large-scale system could be reformulated as follows: distribute evenly the workload to the smallest set of servers operating at optimal or near-optimal energy levels, while observing the Service Level Agreement (SLA) between the CSP and a cloud user. An optimal energy level is one when the performance per Watt of power is maximized.

In order to integrate business requirements and application level needs, in terms of Quality of Service (QoS), cloud service provisioning is regulated by Service Level Agreements (SLAs): contracts between clients and providers that express the price for a service, the QoS levels required during the service provisioning, and the penalties associated with the SLA violations. In such a context, performance evaluation plays a key role allowing system managers to evaluate the effects of different resource management strategies on the data center functioning and to predict the corresponding costs/benefits.

## DISADVANTAGES OF EXISTING SYSTEM:

- On-the-field experiments are mainly focused on the offered QoS, they are based on a black box approach that makes difficult to correlate obtained data to the internal resource management strategies implemented by the system provider.

- Simulation does not allow to conduct comprehensive analyses of the system performance due to the great number of parameters that have to be investigated.

## PROPOSED SYSTEM:

There are three primary contributions of this paper:

- a new model of cloud servers that is based on different operating regimes with various degrees of \energy efficiency" (processing power versus energy consumption);
- a novel algorithm that performs load balancing and application scaling to maximize the number of servers operating in the energy-optimal regime; and
- analysis and comparison of techniques for load balancing and application scaling using three differently-sized clusters and two different average load profiles.

The objective of the algorithms is to ensure that the largest possible number of active servers operate within the boundaries of their respective optimal

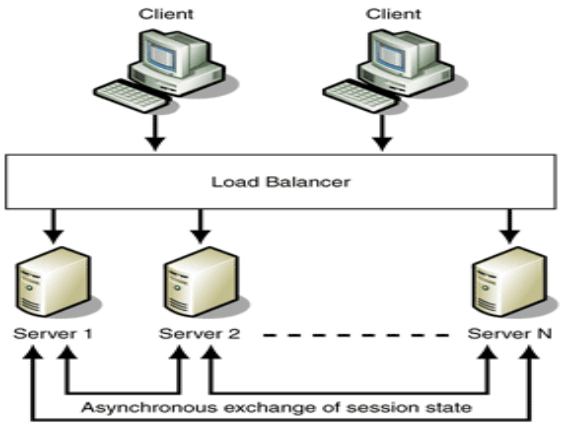
ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

operating regime. The actions implementing this policy are: (a) migrate VMs from a server operating in the undesirable-low regime and then switch the server to a sleep state; (b) switch an idle server to a sleep state and reactivate servers in a sleep state when the cluster load increases; (c) migrate the VMs from an overloaded server, a server operating in the undesirable-high regime with applications predicted to increase their demands for computing in the next reallocation cycles.

## ADVANTAGES OF PROPOSED SYSTEM:

- After load balancing, the number of servers in the optimal regime increases from 0 to about 60% and a fair number of servers are switched to the sleep state.
- There is a balance between computational efficiency and SLA violations; the algorithm can be tuned to maximize computational efficiency or to minimize SLA violations according to the type of workload and the system management policies.

## SYSTEM ARCHITECTURE:




## IMPLEMENTATION MODULES:

- System Model
- Server
- Creating Load
- Energy Aware Load balance

## MODULES DESCRIPTION:

### System Model

In this module, we design the system, such that client makes request to server. Usually, a it is designed with adequate resources in order to satisfy the traffic volume generated by end-users. In general, a wise provisioning of resources can ensure that the input rate is always lower than the service rate. In such a case, the system will be capable to efficiently serve all users' requests. Applications for one instance family have similar profiles, e.g., are CPU-, memory-, or I/O-intensive and run on clusters optimized for that profile; thus, the application interference with one another is minimized. The normalized system performance and the normalized power consumption are different from server to server; yet, warehouse scale computers supporting an instance family use the same processor or family of processors and this reduces the effort to determine the parameters required by our model. In our model the migration decisions are based solely on the vCPU units demanded by an application and the available capacity of the host and of the other servers in the cluster. The model could be extended to take into account not only the processing power, but also the dominant resource for a particular instance family, e.g., memory for R3, storage for I2, GPU for G2 when deciding to migrate a VM. This extension would complicate the model and add additional overhead for monitoring the application behavior

### Server

The term server consolidation is used to describe: switching idle and lightly loaded systems to a sleep state; (2) workload migration to prevent overloading of systems; or (3) any optimization of cloud performance and energy efficiency by redistributing the workload. In this module we design the Server System, where the server processes the client request. Cloud is a large distributed system of servers deployed in multiple data centers across the Internet. The goal of a cloud is to serve content to end-users with high availability and high performance. Cloud serves a large fraction of the Internet content today, including web objects (text, graphics and scripts), downloadable objects (media

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

files, software, documents), applications (e-commerce, portals), live streaming media, on-demand streaming media, and social networks. Besides better performance and availability, cloud also offload the traffic served directly from the content provider's origin infrastructure, resulting in cost savings for the content provider.

### Creating Load

In this module, we create the load to the server. Though, in this paper we focus exclusively on critical conditions where the global resources of the network are close to saturation. This is a realistic assumption since an unusual traffic condition characterized by a high volume of requests, i.e., a flash crowd, can always overfill the available system capacity. In such a situation, the servers are not all overloaded. Rather, we typically have local instability conditions where the input rate is greater than the service rate. In this case, the balancing algorithm helps prevent a local instability condition by redistributing the excess load to less loaded servers.

### Energy Aware Load balance

The objective of the algorithms is to ensure that the largest possible number of active servers operate within the boundaries of their respective optimal operating regime. The actions implementing this policy are: (a) migrate VMs from a server operating in the undesirable-low regime and then switch the server to a sleep state; (b) switch an idle server to a sleep state and reactivate servers in a sleep state when the cluster load increases; (c) migrate the VMs from an overloaded server, a server operating in the undesirable-high regime with applications predicted to increase their demands for computing in the next reallocation cycles. We present a new mechanism for redirecting incoming client requests to the most appropriate server, thus balancing the overall system requests load. Our mechanism leverages local balancing in order to achieve global balancing. This is carried out through a periodic interaction among the system nodes. Depending on the network layers and mechanisms involved in the process, generally request routing techniques can be classified in cloud request routing, transport-layer request routing, and application-layer request routing.

## SCREEN SHOTS

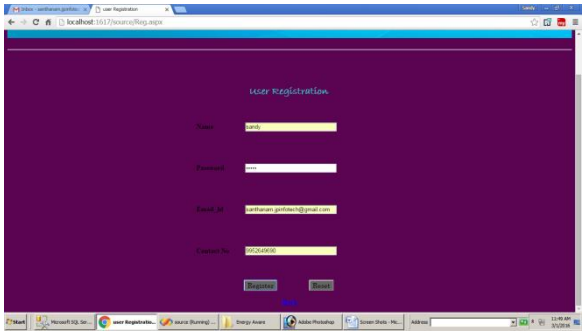
### Admin login:


### Add Server:

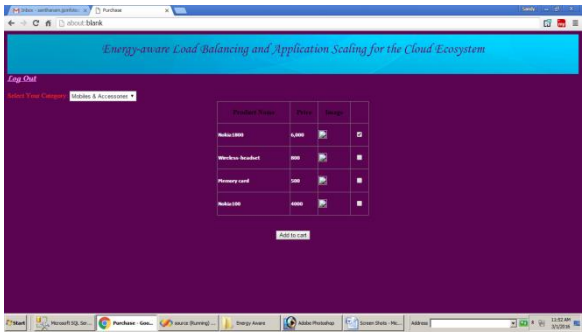

### Server Configuration:

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

**User Registration:**




**User Home:**




## CONCLUSION

The realization that power consumption of cloud computing centers is signi_cant and is expected to increase substantially in the future motivates the interest of the research community in energy-aware resource management and application placement policies and the mechanisms to enforce these policies. Low average server utilization and its impact on the environment make it imperative to devise new energy-aware policies which identify optimal regimes for the cloud servers and, at the same time, prevent SLA violations A quantitative evaluation of an optimization algorithm or an architectural enhancement is a rather intricate and time consuming process; several benchmarks and system configurations are used to gather the data necessary to guide future developments. For example, to evaluate the effects of architectural enhancements supporting Instruction-level or Data-level

Parallelism on the processor performance and their power consumption several benchmarks are used. The results show different numerical outcomes for the individual applications in each benchmark. Similarly, the effects of an energy-aware algorithm depend on the system configuration and on the application and cannot be expressed by a single numerical value. Research on energy-aware resource management in largescale systems often use simulation for a quasi-quantitative and, more often, a qualitative evaluation of optimization algorithms or procedures. As stated in \First, they (WSCs) are a new class of large-scale machines driven by a new and rapidly evolving set of workloads. Their size alone makes them difficult to experiment with, or to simulate efficiently." It is rather diffiult to experiment with the systems discussed in this paper and this is precisely the reason why we choose simulation. The results of the measurements reported in the literature are difficult to relate to one another. For example, the wakeup time of servers in the sleep state and the number of servers in the sleep state are reported for the AutoScale system; yet these configures would be different for another processor, system configuration, and application. We choose computational efficiency, the ratio of the amount of normalized performance to normalized power consumption, as the performance measure of our algorithms. The amount of useful work in a transition processing benchmark can be measured by the number of transactions, but it is more difficult to assess for other types of applications.

## REFERENCES

[1] D. Ardagna, B. Panicucci, M. Trubian, and L. Zhang. \Energy-aware autonomic resource allocation in multitier virtualized environments." IEEE Trans. on Services Computing, 5(1):2{19, 2012.

[2] J. Baliga, R.W.A. Ayre, K. Hinton, and R.S. Tucker \Green cloud computing: balancing energy in processing, storage, and transport." Proc. IEEE, 99(1):149-167, 2011.

[3] L. A. Barroso and U. H□ozle. \The case for energyproportional computing." IEEE Computer, 40(12):33{ 37, 2007.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[4] L. A. Barossso, J. Clidaras, and U.H□ozle. The Data- center as a Computer; an Introduction to the Design of Warehouse-Scale Machines. (Second Edition). Morgan & Claypool, 2013.

[5]A.Beloglazov,R.Buyya\Energye cient resource management in virtualized cloud data centers." Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Comp., 2010.

[6] A. Beloglazov, J. Abawajy, R. Buyya. \Energy-aware resource allocation heuristicsfore cient management of data centers for Cloud computing." Future Generation Computer Systems, 28(5):755-768, 2012.

[7] A. Beloglazov and R. Buyya. \Managing overloaded hosts for dynamic consolidation on virtual machines in cloud centers under quality of service constraints." IEEE Trans. on Parallel and Distributed Systems, 24(7):1366- 1379, 2013.

[8] M. Blackburn and A. Hawkins. \Unused server survey results analysis." www.thegreengrid.org/media/White Papers/Unused%20Server%20Study WP 101910 v1. ashx?lang=en (Accessed on December 6, 2013).

[9] M. Elhawary and Z. J. Haas. \Energy-efficient protocol for cooperative networks." IEEE/ACM Trans. on Net- working, 19(2):561{574, 2011.

[10] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch. \AutoScale: dynamic, robust capacity management for multi-tier data centers." ACM Trans. On Computer Systems, 30(4):1{26, 2012.

[11] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M.Kozuch. \Are sleep states ective in data centers?" Proc. Int. Conf. on Green Comp., pp. 1{10, 2012.

[12] D. Gmach, J. Rolia, L. Cherkasova, G. Belrose, T. Tucricchi, and A. Kemper. \An integrated approach to resource pool management: policies, e_ciency, and quality metrics." Proc. Int. Conf. on Dependable Systems and Networks, pp. 326{335, 2008.

[13] Google. \Google's green computing: e_ciency at scale."http://static.googleusercontent.com/external content/untrusted dlcp/www.google.com/en/us/green/pdfs/google-green-computing.pdf (Accessed on August 29, 2013).

[14] V. Gupta and M. Harchol-Balter. \Self-adaptive admission control policies for resource-sharing systems." Proc. 11th Int. Joint Conf. Measurement and Modeling Computer Systems (SIGMETRICS'09), pp. 311{322, 2009.

[15] K. Hasebe, T. Niwa, A. Sugiki, and K. Kato. \Powersaving in large-scale storage systems with data migration." Proc IEEE 2nd Int. Conf. on Cloud Comp. Technology and Science, pp. 266{273, 2010.