Volume 3, Issue 3, March 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Eco-efficient Approaches to Cloud Computing: A Review

Kamaljit Kaur[*]
CSE Department,
BBSBEC, India

Asmita Pandey
CSE Department,
BBSBEC, India

Abstract— Cloud computing has emerged as a new paradigm of computing and gains increasing attention from both academic and business community. Its utility-based usage model allows users to pay per use, similar to other public utility such as electricity, with relatively low investment on the end devices that access the cloud computing resources. From the environmental perspective this new computing model is already a great improvement since the computing resources are shared among all users and provisioned on-demand. However, data centers hosting Cloud applications consume huge amounts of electrical energy, contributing to high operational costs and carbon footprints to the environment. Therefore, we need Green Cloud computing solutions that can not only minimize operational costs but also reduce the environmental impact. In this paper, various energy efficient methods of cloud computing will be discussed.

Keywords— Cloud computing, Green computing, DVFS, ECTC, MaxUtil, Energy efficient approaches.

I. Introduction

A. *Cloud Computing :*

Cloud computing is an emerging model for distributed utility computing and is being considered as an attractive opportunity for saving energy through central management of computational resources. Clouds[1] aim to power the next generation data centers by architecting them as a network of virtual services (hardware, database, user-interface, application logic) so that users are able to access and deploy applications from anywhere in the world on demand at competitive costs depending on users QoS (Quality of Service) requirements. Developers with innovative ideas for new Internet services are no longer required to make large capital outlays in the hardware and software infrastructures to deploy their services or human expense to operate it. It offers significant benefit to IT companies by freeing them from the low level task of setting up basic hardware (servers) and software infrastructures and thus enabling more focus on innovation and creation of business values.

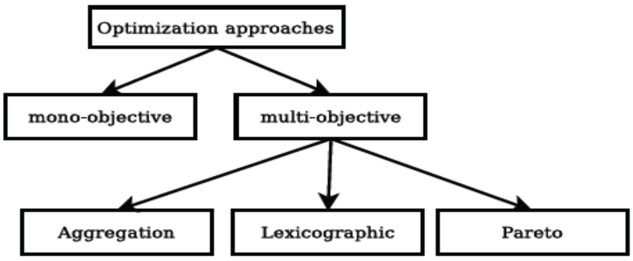
B. *Green Computing :*

Green Computing, or Green IT [2], is the practice of implementing policies and procedures that improve the efficiency of computing resources in such a way as to reduce the energy consumption and environmental impact of their utilization.

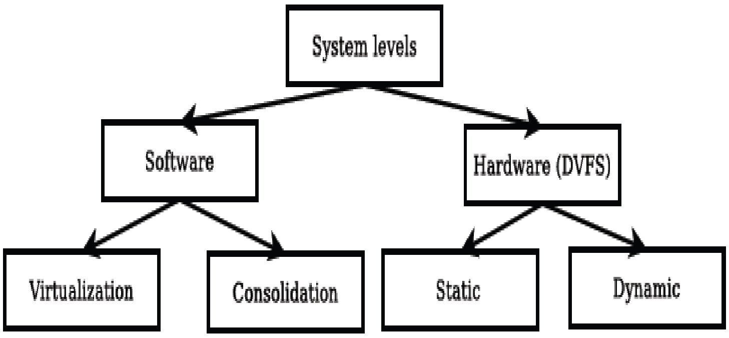
C. *Energy Efficiency :*

Currently, a large number of cloud computing systems waste a tremendous amount of energy and emit a considerable amount of carbon dioxide. Thus, it is necessary to significantly reduce pollution and substantially lower energy usage. The analysis of energy consumption in cloud computing consider both public and private clouds. Cloud computing with green algorithm can enable more energy-efficient use of computing power. To address this problem and drive Green Cloud computing, data center resources need to be managed in an energy-efficient manner. In particular, Cloud resources need to be allocated not only to satisfy Quality of Service (QoS) requirements specified by users via Service Level Agreements (SLAs), but also to reduce energy usage.
The energy aware approaches can be classified into following three types:

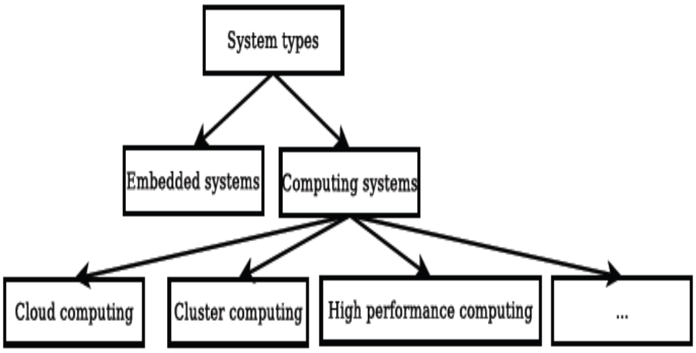
1. **Optimization Approaches:** The first criterion concerns the type of the used optimization. This optimization can be mono-objective or multi-objective.



2. **System levels:** The system level is the second criterion on which methods can be classified. To reduce energy consumption, the methods are based either on software or hardware part of a system. In the hardware level, the techniques used are generally virtualisation and consolidation based approaches. While in the software level, two techniques can be identified to address the task scheduling problem: Static methods where scheduling is done before the execution of a program, and dynamic methods, where the appropriate scheduling is calculated during the execution of a program.



3. **System types:** The third and last criterion used in our classification is the type of system to which the approach is intended to be used. In this criterion, a system can be either a computing or an embedded system. Reducing energy consumption in embedded systems aims to increase the autonomy of devices.



II. RELATED WORK

One of the first works, in which power management has been applied at the data center level, has been done by Pinheiro et al. [3]. In this work the authors have proposed a technique for minimization of power consumption in a heterogeneous cluster of computing nodes serving multiple web-applications. The main technique applied to minimize power consumption is concentrating the workload to the minimum of physical nodes and switching idle nodes off. This approach requires dealing with the power/performance trade-off, as performance of applications can be degraded due to the workload consolidation. Requirements to the throughput and execution time of applications are defined in SLAs to ensure reliable QoS. The proposed algorithm periodically monitors the load of resources (CPU, disk storage and network interface) and makes decisions on switching nodes on/off to minimize the overall power consumption, while providing the expected performance. The actual load balancing is not handled by the system and has to be managed by the applications. The algorithm runs on a master node, which creates a Single Point of Failure (SPF) and may become a performance bottleneck in a large system. In addition, the authors have pointed out that the reconfiguration operations are time-consuming, and the algorithm adds or removes only one node at a time, which may also be a reason for slow reaction in large-scale environments. The proposed approach can be applied to multi-application mixed-workload environments with fixed SLAs Chase et al. [4] have considered the problem of energy-efficient management of homogeneous resources in Internet hosting centers. The main challenge is to determine the resource demand of each application at its current request load level and to allocate resources in the most efficient way. To deal with this problem the authors have applied an economic framework: services ‘‘bid’’ for resources in terms of volume and quality. This enables negotiation of the SLAs according to the available budget and current QoS requirements, i.e. balancing the cost of resource usage (energy cost) and the benefit gained due to the usage of this resource. The system maintains an active set of servers selected to serve requests for each service. The network switches are dynamically reconfigured to change the active set of servers when necessary. Energy consumption is reduced by switching idle servers to power saving modes (e.g. sleep, hibernation). The system is targeted at the web workload, which leads to a ‘‘noise’’ in the load data. The authors have addressed this problem by applying the statistical ‘‘flip-flop’’ filter, which reduces the number of unproductive reallocations and leads to a more stable and efficient control. The proposed approach is suitable for multi-application environments with variable SLAs and has created a foundation for numerous studies on power-efficient resource allocation at the data center level. However, in contrast to [3], the system deals only with the management of the CPU, but does not consider other system resources. The latency due to switching nodes on/off also is not taken into account. The authors have noted that the management algorithm is fast when the workload is stable, but turns out to be relatively expensive during significant changes in the workload. Moreover, likewise [3], diverse software configurations are not handled, which can be addressed by applying the virtualization technology.

Elnozahy et al. [5] have investigated the problem of power-efficient resource management in a single web-application environment with fixed SLAs (response time) and load balancing handled by the application. As in [4], two power saving techniques are applied: switching power of computing nodes on/off and Dynamic Voltage and Frequency Scaling (DVFS). The main idea of the policy is to estimate the total CPU frequency required to provide the necessary response time, determine the optimal number of physical nodes and set the proportional frequency to all the nodes. However, the transition time for switching the power of a node is not considered. Only a single application is assumed to be run in the system and, like in [3], the load balancing is supposed to be handled by an external system. The algorithm is centralized that creates an SPF and reduces the scalability. Despite the variable nature of the workload, unlike [4], the resource usage data are not approximated, which results in potentially inefficient decisions due to fluctuations.

Nathuji and Schwan [6] have studied power management techniques in the context of virtualized data centers, which has not been done before. Besides hardware scaling and VMs consolidation,the authors have introduced and applied a new power management technique called ‘‘soft resource scaling’’. The idea is to emulate hardware scaling by providing less resource time for a VM using the Virtual Machine Monitor’s (VMM) scheduling capability. The authors found that a combination of ‘‘hard’’ and ‘‘soft’’ scaling may provide higher power savings due to the limited number of hardware scaling states. The authors have proposed an architecture where the resource management is divided into local and global policies. At the local level the system leverages the guest OS’s power management strategies. However, such management may appear to be inefficient, as the guest OS may be legacy or power-unaware.

Raghavendra et al. [7] have investigated the problem of power management for a data center environment by combining and coordinating five diverse power management policies. The authors explored the problem in terms of control theory and applied a feedback control loop to coordinate the controllers’ actions. It is claimed that, similarly to [6], the approach is independent of the workload type. Like most of the previous works, the system deals only with the CPU management. The authors have pointed out an interesting outcome of the experiments: the actual power savings can vary depending on the workload, but ‘‘the benefits from coordination are qualitatively similar for all classes of workloads’’. However, the system fails to support strict SLAs as well as variable SLAs for different applications. This results in the suitability for enterprise environments, but not for Cloud computing providers, where more comprehensive support for SLAs is essential.

Kusic et al. [8] have defined the problem of power management in virtualized heterogeneous environments as a sequential optimization and addressed it using Limited Lookahead Control (LLC). The objective is to maximize the resource provider’s profit by minimizing both power consumption and SLA violation. Kalman filter is applied to estimate the number of future requests to predict the future state of the system and perform necessary reallocations. However, in contrast to heuristic based approaches, the proposed model requires simulation-based learning for the application specific adjustments. Moreover, due to the complexity of the model the execution time of the optimization controller reaches 30 min even for 15 nodes, which is not suitable for large-scale real-world systems.

Srikantaiah et al. [9] have studied the problem of request scheduling for multi-tiered web-applications in virtualized heterogeneous systems to minimize energy consumption, while meeting performance requirements. The authors have investigated the effect of performance degradation due to high utilization of different resources when the workload is consolidated. They have found that the energy consumption per transaction results in a ‘‘U’’-shaped curve, and it is possible to determine the optimal utilization point. To handle the optimization over multiple resources, the authors have proposed a heuristic for the multidimensional bin packing problem as an algorithm for the workload consolidation. However, the proposed approach is workload type and application dependent, whereas our algorithms are independent of the workload type, and thus are suitable for a generic Cloud environment.

Cardosa et al. [10] 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 VMM, which represent minimum, maximum and proportion of the CPU allocated to VMs sharing the same resource. Similarly to [7], 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) and no other resources except for the CPU are considered during the VM reallocation.

Verma et al. [11] 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. Like in [9], the authors have applied a heuristic for the bin packing problem with variable bin sizes and costs. Similarly to [6], live migration of VMs is used to achieve a new placement at each time frame. The proposed algorithms, on the contrary to our approach, do not handle strict SLA requirements: SLAs can be violated due to variability of the workload. Gandhi et al. [12] have considered the problem of allocating an available power budget among servers in a virtualized heterogeneous server farm, while minimizing the mean response time. To investigate the effect of different factors on mean response time, a queuing theoretic model has been introduced, which allows the prediction of the mean response time as a function of the power-to-frequency relationship, arrival rate, peak power budget, etc. The model is used to determine the optimal power allocation for every configuration of the above factors.

Gupta et al. [13] have suggested putting network interfaces, links, switches and routers into sleep modes when they are idle in order to save the energy consumed by the Internet backbone and consumers. Based on the foundation laid by Gupta et al. [13], a number of research works have been done on the energy-efficient traffic routing by ISPs and applying sleep modes and performance scaling of network devices [14,15]. Chiaraviglio and Matta [16] have proposed cooperation between ISPs and content providers that allows the achievement of an efficient simultaneous allocation of compute resources and network paths that minimizes energy consumption under performance constraints. Koseoglu and Karasan [17] have applied a similar approach of joint allocation of computational resources and network paths to Grid environments based on the optical burst switching technology with the objective of minimizing job completion times. Tomas et al. [18] have investigated the problem of scheduling Message Passing Interface (MPI) jobs in Grids considering network data transfers satisfying the QoS requirements.

Dodonov and de Mello [19] have proposed an approach to scheduling distributed applications in Grids based on predictions of communication events. They have proposed the migration of communicating processes if the migration cost is lower than the cost of the predicted communication with the objective of minimizing the total execution time. They have shown that the approach can be effectively applied in Grids; however, it is not viable for virtualized data centers, as the VM migration cost is higher than the process migration cost. Gyarmati and Trinh [20] have investigated the energy consumption implications of data centers' network architectures. However, optimization of network architectures can be applied only at the data center design time and cannot be applied dynamically. Guo et al. [21] have proposed and implemented a virtual cluster management system that allocates the resources in a way satisfying bandwidth guarantees. The allocation is determined by a heuristic that minimizes the total bandwidth utilized. The VM allocation is adapted when some of the VMs are de-allocated. However, the VM allocation is not dynamically adapted depending on the current network load. Moreover, the approach does not explicitly minimize energy consumption by the network. Rodero-Merino et al. [22] have proposed an additional layer of infrastructure management in Clouds with the ability to automatically deploy services with multi-VM configurations. The proposed infrastructure management system applies the specified rules for scaling VM configurations in and out. However, the system does not optimize the network communication between VMs. Calheiros et al. [23] have investigated the problem of mapping VMs on physical nodes optimizing network communication between VMs; however, the problem has not been explored in the context of the optimization of energy consumption.

Young Choon Lee et al. in [24], present two energy-conscious task consolidation heuristics *ECTC* and *MaxUtil*. They are in fact described side by side since they share several common features with the main difference being whether energy consumption is taken into account explicitly or implicitly. Both of them aim to maximize resource utilization and explicitly take into account both active and idle energy consumption. These heuristics assign each task to the resource on which the energy consumption for executing the task is explicitly or implicitly minimized without the performance degradation of that task. Based on their experimental results, these heuristics demonstrate their promising energy-saving capability.

Beloglazov et al. [25] discuss another benchmark VM selection policy that is a VM migration aware policy called Single Threshold (ST). It is based on the idea of setting the upper utilization threshold for hosts and placing VMs, while keeping the total utilization of CPU below this threshold. At each time frame all VMs are reallocated using the MBFD algorithm with additional condition of keeping the upper utilization threshold not violated. The results showed that with the growth of the utilization threshold energy consumption decreases, whereas the percentage of SLA violations increases. This is due to the fact that a higher utilization threshold allows more aggressive consolidation of VMs by the cost of the increased risk of SLA violations.

Laszewski et al. [26] focus on scheduling virtual machines in a computer cluster to reduce power consumption via the technique of Dynamic Voltage Frequency Scaling (DVFS). Specifically, they presented the design and implementation of an efficient scheduling algorithm to allocate virtual machines in a DVFS-enabled cluster by dynamically scaling the supplied voltages. The algorithm has been studied via simulation and implementation in a multi-core cluster.

III. CONCLUSIONS

There are several methods of achieving energy efficiency in cloud computing systems. The latest trend is to implement efficient resource allocation algorithms in order to reduce the energy consumption. Maximizing the resource utilization results in reduced energy consumption and increased SLA violations. So, it is a complex problem to achieve a balance between energy efficiency and QoS requirements.

REFERENCES

[1] A. Weiss. Computing in the clouds. NetWorker, 11(4):16–25, Dec. 2007.

[2] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. D. Rose, and R. Buyya, CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms, Software: Practice and Experience, Wiley Press, New York, USA, 2010.

[3] E. Pinheiro, R. Bianchini, E.V. Carrera, T. Heath, Load balancing and unbalancing for power and performance in cluster-based systems, in: Proceedings of the Workshop on Compilers and Operating Systems for Low Power, 2001, pp. 182–195.

[4] J.S. Chase, D.C. Anderson, P.N. Thakar, A.M. Vahdat, R.P. Doyle, Managing energy and server resources in hosting centers, in: Proceedings of the 18th ACM Symposium on Operating Systems Principles, ACM, New York, NY, USA, 2001, pp. 103–116.

[5] E. Elnozahy, M. Kistler, R. Rajamony, Energy-efficient server clusters, Power- Aware Computer Systems (2003) 179–197.

[6] R. Nathuji, K. Schwan, Virtualpower: coordinated power management in virtualized enterprise systems, ACM SIGOPS Operating Systems Review 41 (6) (2007) 265–278.

[7] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, X. Zhu, No ‘‘power’’ struggles: coordinated multi-level power management for the data center, SIGARCH Computer Architecture News 36 (1) (2008) 48–59.

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

[9] S. Srikantaiah, A. Kansal, F. Zhao, Energy aware consolidation for cloud computing, Cluster Computing 12 (2009) 1–15.

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

[11] 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.

[12] A. Gandhi, M. Harchol-Balter, R. Das, C. Lefurgy, Optimal power allocation in server farms, in: Proceedings of the 11th International Joint Conference on Measurement and Modeling of Computer Systems, ACM, New York, NY, USA, 2009, pp. 157–168.

[13] M. Gupta, S. Singh, Greening of the internet, in: Proceedings of the ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM 2003, New York, NY, USA, 2003, pp. 19–26.

[14] N. Vasic, D. Kostic, Energy-aware traffic engineering, in: Proceedings of the 1st ACM International Conference on Energy-Efficient Computing and Networking, e-Energy 2010, Passau, Germany, 2010, pp. 169–178.

[15] C. Panarello, A. Lombardo, G. Schembra, L. Chiaraviglio, M. Mellia, Energy saving and network performance: a trade-off approach, in: Proceedings of the 1st ACM International Conference on Energy-Efficient Computing and Networking, e-Energy 2010, Passau, Germany, 2010, pp. 41–50.

[16] L. Chiaraviglio, I. Matta, GreenCoop: cooperative green routing with energy efficient servers, in: Proceedings of the 1st ACM International Conference on Energy-Efficient Computing and Networking, e-Energy 2010, Passau, Germany, 2010, pp. 191–194.

[17] M. Koseoglu, E. Karasan, Joint resource and network scheduling with adaptive offset determination for optical burst switched grids, Future Generation Computer Systems 26 (4) (2010) 576–589.

[18] L. Tomas, A. Caminero, C. Carrion, B. Caminero, Network-aware metascheduling in advance with autonomous self-tuning system, Future Generation Computer Systems 27 (5) (2010) 486–497.

[19] E. Dodonov, R. de Mell, A novel approach for distributed application scheduling based on prediction of communication events, Future Generation Computer Systems 26 (5) (2010) 740–752.

[20] L. Gyarmati, T. Trinh,Howcan architecture help to reduce energy consumption in data center networking? in: Proceedings of the 1st ACM International Conference on Energy-Efficient Computing and Networking, e-Energy 2010, Passau, Germany, 2010, pp. 183–186.

[21] C. Guo, G. Lu, H. Wang, S. Yang, C. Kong, P. Sun, W. Wu, Y. Zhang, Secondnet: a data center network virtualization architecture with bandwidth guarantees, in: Proceedings of the 6th International Conference on Emerging Networking Experiments and Technologies, CoNEXT 2010, Philadelphia, USA, 2010.

[22] L. Rodero-Merino, L. Vaquero, V. Gil, F. Galan, J. Fontan, R. Montero, I. Llorente, From infrastructure delivery to service management in clouds, Future Generation Computer Systems 26 (8) (2010) 1226–1240.

[23] R.N. Calheiros, R. Buyya, C.A.F.D. Rose, A heuristic for mapping virtual machines and links in emulation testbeds, in: Proceedings of the 38th International Conference on Parallel Processing, Vienna, Austria, 2009.

[24] Young Choon Lee, Albert Y. Zomaya: Energy efficient utilization of resources in cloud computing systems, Springer, J Supercomput (2012) 60:268–280

[25] Anton Beloglazov, Jemal Abawajyb, Rajkumar Buyyaa, Energy-aware resource allocation heuristics for efficient management of datacenters for Cloud computing, Future Generation Computer Systems 28 (2012) 755–768

[26] Von Laszewski, G.L. Wang, A.J. Younge, X.He, Power-Aware Scheduling of Virtual Machines in DVFS-enabled Clusters ,published in Proceedings of IEEE International Conference on Cluster Computing and Workshops, 2009. CLUSTER '09., New Orleans, LA, pages 1-10.