
Reducing Energy Consumption In Cloud Data Centers Via Multi-Sleep Mode Server Scheduling

P Raveena ¹, Dr K Arun Kumar²

¹PG Scholar , Department of Computer Science and Engineering , Malla Reddy Engineering College(A), Maisammaguda,Secunderabad-500100, Telangana,India.

²Associate Professor, Department of Computer Science and Engineering ,Malla Reddy Engineering College(A), Maisammaguda,Secunderabad-500100, Telangana, India.

Email: ¹pulusugantiraveena@gmail.com, ²kandruarunpg@gmail.com

In a cloud data center, servers are always over-provisioned in an active state to meet the peak demand of requests, wasting a large amount of energy as a result. One of the options to reduce the power consumption of data centers is to reduce the number of idle servers, or to switch idle servers into low-power sleep states. However, the servers cannot process the requests immediately when transitioning to an active state. There are delays and extra power consumption during the transition. In this paper, we consider using state-of-the-art servers with multi-sleep modes. The sleep modes with smaller transition delays usually consume more power when sleeping. Given the arrival of incoming requests, our goal is to minimize the energy consumption of a cloud data center by the scheduling of servers with multi-sleep modes. We formulate this problem as an integer linear programming (ILP) problem during the whole period of time with millions of decision variables. To solve this problem, we divide it into sub-problems with smaller periods while ensuring the feasibility and transition continuity for each sub-problem through a Backtrack-and-Update technique. We also consider using DVFS to adjust the frequency of active servers, so that the requests can be processed with the least power. Our simulations are based on traces from real world. Experiments show that our method can significantly reduce the power consumption for a cloud data center.

Key words: AES ,integer linear programming ,Cloud, Energy , DVFS,VMS, Servers.

I INTRODUCTION

In recent years, cloud data centers are expanding rapidly to meet the ever increasing demand of computing capacity. It is the powerful servers of the data centers that consume a huge amount of energy. According to a report, data centers consume about 1.3% of the worldwide electricity, which is expected to reach 8% in 2020 [1]. Meanwhile, much of the energy is wasted, because servers are busy only 10% 30% of the time on average, with most time in idle state. What's worse, a server can even consume 60% or more of its peak power when in idleness [2]. To handle the possible peak demand of user requests, servers are always over-provisioned, wasting a lot of energy as a result.

Therefore, there is an urgent need to enhance energy efficiency for cloud data centers. The existing work has mainly focused on dynamic voltage frequency scaling (DVFS) and dynamic power management (DPM). The former is to adjust the voltage/frequency of CPU power according to the demand of computing capacity, while the latter reduces the total energy by putting servers into sleep states or turning off idle servers. However, a difficult issue is that the servers cannot process the incoming requests immediately when transiting to active state. There are delays and extra power consumption during the transitions, which have been ignored in the existing work. Besides, modern servers are usually designed with several sleep states, and the sleep states with smaller transition delays consume more power when sleeping. In this paper, we study the issue of minimizing energy consumption of a data center by scheduling servers in multi-sleep modes and at different frequency levels to reduce the total energy of active servers. That is, given the arrival of user requests, schedule the servers (to active state with different frequencies or to different sleep states), such that the total energy consumption of the data center can be minimized while satisfying the QoS requirement. The scheduling algorithm will determine:

- 1) how many of the active servers should be switched into which sleep state in each timeslot
- 2) how many of the sleeping servers in sleep states should be woken up in each timeslot;
- 3) What frequency levels should the active servers be set to in each timeslot.

The scheduling period of our problem consists of T small timeslots. We solve the problem in two steps. In the first step, we aim to minimize the total number of active servers to meet the QoS requirement by assuming that all servers run at the highest frequency. The problem is formulated as a constraint optimization problem with millions of decision variables due to the large number of timeslots. It is not feasible to solve the problem of such a large size using existing methods. We group multiple timeslots into a segment with equal length, and formulate the scheduling in each segment independently as an integer linear programming (ILP) sub-problem. By using Cplex to solve each sub-problem, the optimal solution can be obtained for each segment. However, the scheduling of the current segment doesn't consider the arrival of the requests in the next segment. It may lead to the situation that some servers are put into sleep at the end of this segment, but cannot be woken up immediately to cope with request burst at the beginning of the next segment. We propose a Backtrack-and-Update technique to solve this issue. In the second step, we make scaling of the frequency levels of the active servers, so that the requests can be processed with the least necessary power. In each timeslot, this problem can also be formulated into an independent ILP problem of a small size that the optimal solution can be obtained. Our simulations are based on traces from real world. Experiments show that our method can significantly reduce the total energy consumption for a cloud data center. The rest of this paper are organized as follows. In Section 2, the modeling and formulation of our problem will be given in detail. Section 3 gives the Backtrack-and-Update algorithm as solution. In Section 4, we set up the experiments and make evaluations. Section 5 reviews the related work. Finally, Section 6 concludes this paper.

II LITERATURE SURVEY

Are sleep states effective in data centers

While sleep states have existed for mobile devices and workstations for some time, these sleep states have not been incorporated into most of the servers in today's data centers. High setup times make data center administrators fearful of any form of dynamic power management, whereby servers are suspended or shut down when load drops. This general reluctance has stalled research into whether there might be some feasible sleep state (with sufficiently low setup overhead and/or sufficiently low power) that would actually be beneficial in data centers. This paper investigates the regime of sleep states that would be advantageous in data centers.

We consider the benefits of sleep states across three orthogonal dimensions:

- (i) the variability in the workload trace,
- (ii) the type of dynamic power management policy employed, and
- (iii) the size of the data center. Our implementation results on a 24-server multi-tier testbed indicate that under many traces, sleep states greatly enhance dynamic power management.

In fact, given the right sleep states, even a naïve policy that simply tries to match capacity with demand, can be very effective. By contrast, we characterize certain types of traces for which even the "best" sleep state under consideration is ineffective. Our simulation results suggest that sleep states are even more beneficial for larger data centers.

Greenware: Greening cloudscale data centers to maximize the use of renewable energy

To reduce the negative environmental implications (e.g., CO₂ emission and global warming) caused by the rapidly increasing energy consumption, many Internet service operators have started taking various initiatives to operate their cloud-scale data centers with renewable energy. Unfortunately, due to the intermittent nature of renewable energy sources such as wind turbines and solar panels, currently renewable energy is often more expensive than brown energy that is produced with conventional fossil-based fuel. As a result, utilizing renewable energy may impose a considerable pressure on the sometimes stringent operation budgets of Internet service operators. Therefore, two key questions faced by many cloud-service operators are

- 1) How to dynamically distribute service requests among data centers in different geographical locations, based on the local weather conditions, to maximize the use of renewable energy, and
- 2) How to do that within their allowed operation budgets. In this paper, we propose GreenWare, a novel middleware system that conducts dynamic request dispatching to maximize the percentage of renewable energy used to power a network of distributed data centers, subject to the desired cost budget of the Internet service operator.

Our solution first explicitly models the intermittent generation of renewable energy, e.g., wind power and solar power, with respect to varying weather conditions in the geographical location of each data center. We then formulate the core objective of GreenWare as a constrained optimization problem and propose an efficient request dispatching algorithm based on linear-fractional programming (LFP). We evaluate GreenWare with real-world weather, electricity price, and workload traces. Our experimental results show that GreenWare can significantly increase the use of renewable energy in cloud-scale data centers without violating the desired cost budget, despite the intermittent supplies of renewable energy in different locations and time-varying electricity prices and workloads.

Experience with using the parallel workloads archive

Science is based upon observation. The scientific study of complex computer systems should therefore be based on observation of how they are used in practice, as opposed to how they are assumed to be used or how they were designed to be used. In particular, detailed workload logs from real computer systems are invaluable for research on performance evaluation and for designing new systems. Regrettably, workload data may suffer from quality issues that might distort the study results, just as scientific observations in other fields may suffer from measurement errors. The cumulative experience with the Parallel Workloads Archive, a repository of job-level usage data from large-scale parallel supercomputers, clusters, and grids, has exposed many such issues. Importantly, these issues were not anticipated when the data was collected, and uncovering them was not trivial. As the data in this archive is used in hundreds of studies, it is necessary to describe and debate procedures that may be used to improve its data quality. Specifically, we consider issues like missing data, inconsistent data, erroneous data, system configuration changes during the logging period, and unrepresentative user behavior. Some of these may be countered by filtering out the problematic data items. In other cases, being cognizant of the problems may affect the decision of which datasets to use. While grounded in the specific domain of parallel jobs, our findings and suggested procedures can also inform similar situations in other domains

Towards optimal electric demand management for internet data centers

Electricity cost is becoming a major portion of Internet data center (IDC)'s operation cost and large-scale IDCs are becoming important consumers of regional electricity markets. IDC's energy efficiency is gaining more attention by data center operators and electricity market operators. Effective IDC electric demand management solutions are eagerly sought by all stakeholders. In this paper, a mixed-integer programming model based IDC electric demand management solution is proposed, which integrates both the impacts of locational marginal electricity prices and power management capability of IDC itself. Dynamic voltage/frequency scaling of individual server, cluster server ON/OFF scheduling, and dynamic workload dispatching are optimized while complying with all the IDC system-wide and individual heterogeneous servers' operation constraints according to the IDC applications' temporal variant workload. Reduced electricity cost can be achieved together with guaranteed QoS requirement and reliability consideration by using the proposed model. World Cup '98 data is utilized to evaluate the effectiveness of the proposed solution. According to the experimental evaluation, electricity cost could be cut by more than 20% in a peak workload period and by more than 80% in a light workload period. Besides, more than 6% electricity cost could be cut by considering the impact of electricity price difference. Experimental results also reveal that higher QoS requirement and reliability consideration could result in higher electricity cost.

In a cloud data center, servers are always over-provisioned in an active state to meet the peak demand of requests, wasting a large amount of energy as a result. One of the options to reduce the power consumption of data centers is to reduce the number of idle servers, or to switch idle servers into low-power sleep states. However, the servers cannot process the requests immediately when transiting to an active state. There are delays and extra power consumption during the transition.

In the existing system, DVFS mechanism scales the CPU chipset power through adjusting the voltage and frequency of CPU. That is, the processing capacity varies with different power levels. Gandhi et al. in [17] combine DFS (Dynamic Frequency Scaling and DVFS to optimize the power allocation in server farm to minimize the response time within a fixed peak power budget. Gerards et al. in [18] try to minimize energy cost through global DVFS on multi-core processors platform while considering the precedence constraint in task scheduling. employ a DVS (Dynamic Voltage Scaling) and node On/Off method to reduce the aggregate power consumption of cluster during periods of reduced workload. They also use both DVS and requests batching mechanisms to reduce processor energy over a wide range of workload intensities .build power models to estimate the energy consumption of user applications under different DVFS policies.

III PROPOSED MODEL

In this paper, we consider using stateof- the-art servers with multi-sleep modes. The sleep modes with smaller transition delays usually consume more power when sleeping. Given the arrival of incoming requests, our goal is to minimize the energy consumption of a cloud data center by the scheduling of servers with multi-sleep modes. We formulate this problem as an integer linear programming (ILP) problem during the whole period of time with millions of decision variables. To solve this problem, we divide it into sub-problems with smaller periods while ensuring the feasibility and transition continuity for each sub-problem through a Backtrack-and-Update technique. We also consider using DVFS to adjust the frequency of active servers, so that the requests can be processed with the least power. Our simulations are based on traces from real world. Experiments show that our method can significantly reduce the power consumption for a cloud data center.

In the proposed system, the system studies the issue of minimizing energy consumption of a data center by scheduling servers in multi sleep modes and at different frequency levels to reduce the total energy of active servers. That is, given the arrival of user requests, schedule the servers (to active state with different frequencies or to different sleep states), such that the total energy consumption of the data center can be minimized while satisfying the QoS requirement.

The scheduling algorithm will determine:

- 1) how many of the active servers should be switched into which sleep state in each timeslot;
- 2) how many of the sleeping servers in sleep states should be woken up in each timeslot;
- 3) What frequency levels should the active servers be set to in each timeslot.

IV METHODOLOGY:

In general, there are three ways to green cloud data centers: dynamic voltage frequency scaling, dynamic power management, and the scheduling using renewable energy.

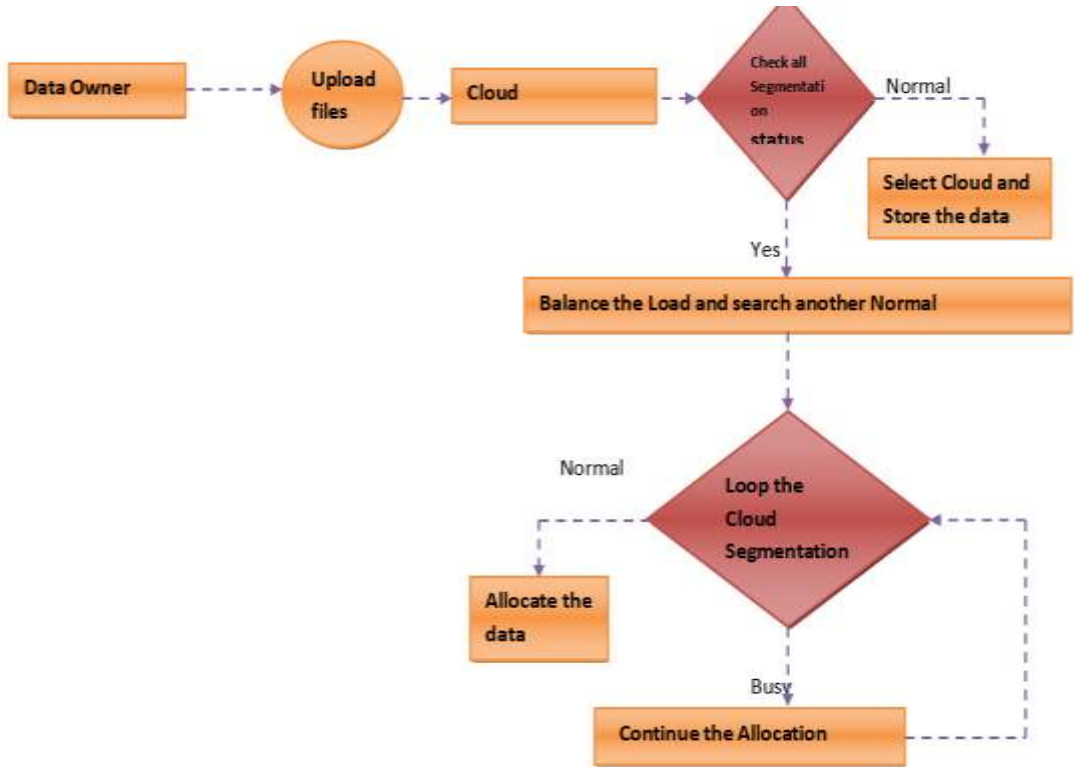


Figure 1: Allocating the memory to the data at Cloud Data Centers

Dynamic Power Management (DPM)

In contrast, the DPM scheme can power down all the components of servers so that the total energy of servers can further be reduced. The commonly used policies of DPM are to power off or switch the idle servers into sleep states, and then wake them up when necessary. reduce the energy consumption of servers through consolidation of virtual machines, while considering the transition delays and energy cost during transitions. They assume a transition only happen within scheduling timeslot, which is impractical in reality. propose a right-sizing method to decide the minimum number of servers that should be online to serve the requests so as to reduce the total energy of servers. Gandhi propose to wait a period of time before turning on servers through a distributed robust auto scaling (DRAS) policy for computing intensive server farms. SoftReactive tries to keep the right number of servers in idle by combination of timer and index-based routing . SoftReactive is a special case for AutoScale that also sets timer to avoid the mistake of turning off a server just as a new arrival comes in. The latter uses a capacity inference algorithm to determine the right number of idle servers using the control knob that can be tracked in Apache and load balancer to scale the number of servers periodically, lacking theoretical support. reduce the power consumption of server farm by formulating the problem as a Markov decision process using sleeping modes. Pinheiro et al. intake load balancing into account when reducing energy through dynamically turning on and off servers, but frequent turning on/off servers is harmful to the hardware and may reduce

lifetime of servers. Meisner et al. in [1] propose an energy-conservation approach called PowerNap, which eliminates idle server power by quickly transitioning in and out of an ultra-low power state, but consider only one sleeping mode. Heath et al. in [2] develop a model-based cooperative Web server to minimize energy consumption for heterogeneous clusters. Chen et al. in [3] propose three strategies based on steady state queuing analysis, feedback control theory, and a hybrid mechanism to minimize operational costs while meeting the SLAs. Liu et al. in [4] present an analytical framework for characterizing and optimizing the power-performance tradeoff in SaaS cloud platform, and the energy saving is achieved by the scheduling of idle/busy states of virtual machines. Matthias et al. in [5] create a Markov model as an abstraction of queuing systems with sleep modes and add delayed activation and deactivation, so as to reduce the total server power. Raj et al. in [6] propose a simple method to reduce the total energy using only one type of sleep mode in combination with shutdown to save energy. However, using only one sleep mode is not enough to reduce total energy when making quick responses.

Green Scheduling Using Renewable Energy

There have been studies on green scheduling using renewable energy for cloud data center. [7] propose a carbon-aware scheduling framework that makes online decisions on geographical load balancing, capacity right-sizing, and server speed scaling using Lyapunov optimization techniques. In our previous work, we try to minimize the total carbon emissions under budget of energy cost when scheduling requests, servers with On/Off switching, and the usage of renewable energy among the data centers in geodistributed locations [8]. To further green cloud data centers, we consider using ESDs (energy storage devices) to store different types of energy with fluctuating prices, so that both energy cost and total carbon emissions can be greatly reduced. To leverage the climate advantages of different locations, we propose a green plan that optimally deploys wind turbines, solar panels and ESDs for future green cloud data centers [9]. In reality, the data centers may have their own wind or solar farms. The self-generated green energy can be used to power data centers directly or sold back to the power grid, so that the data centers can be greened with lower cost [10]. However, the scheduling granularity in these work is usually very large, which is set to be from 10 minutes to 1 hour. Therefore, there is still large room to improve the energy efficiency through scheduling of servers with multi-sleep modes.

As far as we know, there is not too much research about scheduling of servers with multi-sleep modes in energy saving. The possibility of using multiple sleep modes was first studied by Horvath in [11]. However, the importance of the transition power and delay of different sleep modes is not well considered. After that, [12] propose a self-adaptive management of the sleep depths to reduce the total energy of servers. Unfortunately, the transition delay and power during sleep-down process have been ignored in their model. Different from existing work, we focus on minimizing the total power consumption of the cloud data center by switching servers between active state and multiple sleep states according to the varying incoming requests. The novelty of our problem is that we take into account of transition delays and power, and that consumed under different sleep modes in our scheduling, which greatly affect the decisions in energy saving for cloud data center.

AES Overview

AES is a symmetric encryption algorithm that uses the same key for both encryption and decryption. It supports key sizes of 128, 192, and 256 bits, with AES-256 being the most commonly used for high-security applications. AES is fast and efficient for encrypting large amounts of data, which makes it ideal for cloud environments.

AES in Cloud Computing

- **Data Encryption at Rest:** Cloud service providers (CSPs) use AES to encrypt stored data (at rest), such as files, databases, and backups. Even if an attacker gains access to the storage infrastructure, the data remains unreadable without the decryption key.
- **Data Encryption in Transit:** Data moving between a user's device and the cloud is often encrypted using protocols like TLS, which in turn use AES to protect the communication. This ensures that data cannot be intercepted or tampered with.
- **Data Encryption during Processing:** Some cloud environments employ techniques like homomorphic encryption or secure enclaves to perform computations on encrypted data without decrypting it. AES is often used in such setups as part of broader encryption strategies.

3. Key Management in the Cloud

- **Customer-Managed Keys:** Cloud providers like AWS, Azure, and Google Cloud offer key management systems (KMS) that allow customers to manage their own AES encryption keys. This gives users more control over who can access their encrypted data.
- **Provider-Managed Keys:** Cloud providers can also manage the encryption keys on behalf of customers, simplifying key management but potentially reducing direct control.

AES Performance in the Cloud

AES encryption is efficient and scalable, which is crucial in cloud environments where large volumes of data are processed. Many cloud providers offer hardware-based acceleration (e.g., Intel AES-NI) to optimize AES encryption and decryption processes, reducing latency and improving performance.

Compliance and AES

AES is widely adopted due to its strong security properties and compliance with various security standards, such as:

- **FIPS 140-2:** For government use
- **PCI-DSS:** For handling payment card data
- **HIPAA:** For healthcare data protection

V SCREEN SHOTS

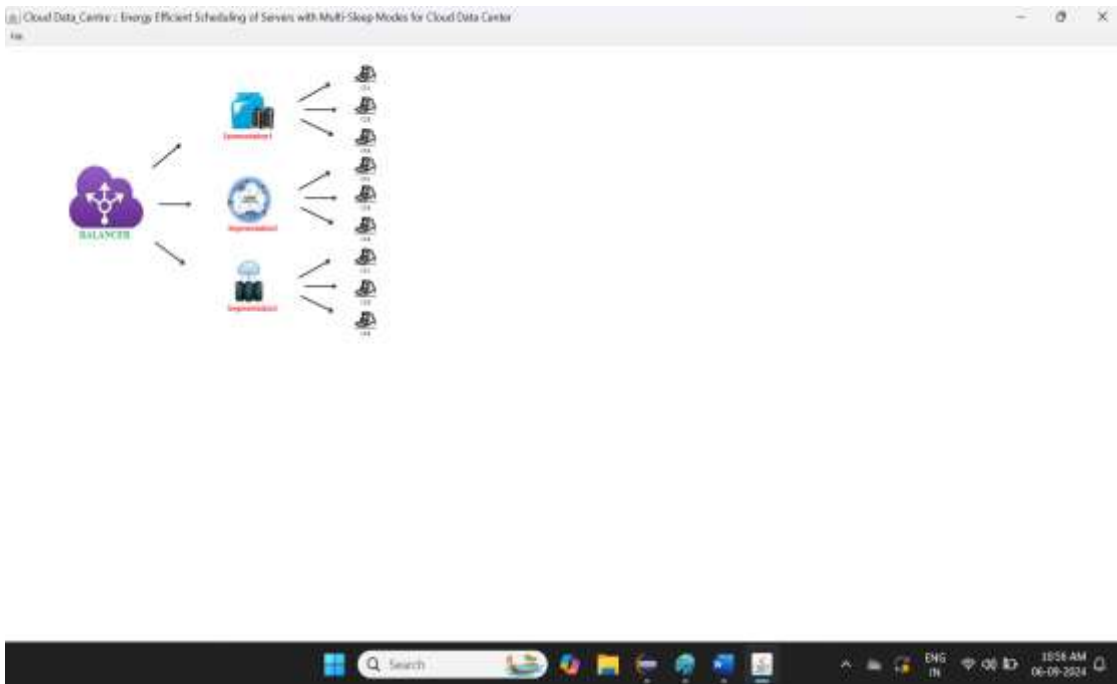


Figure 2: Overview of cloud energy Efficient

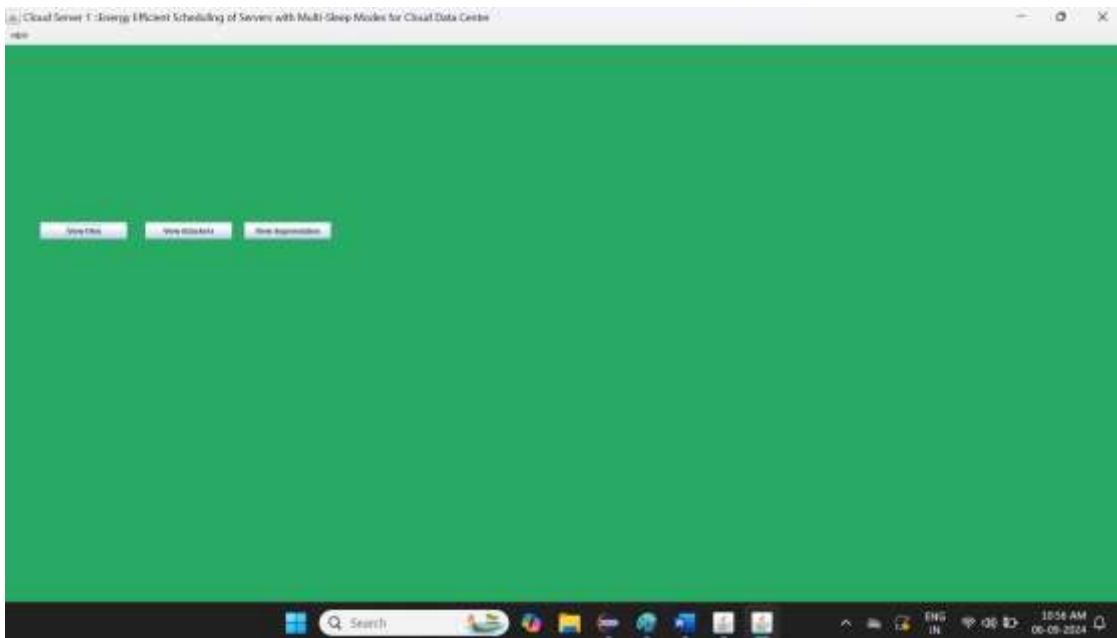


Figure 3: Different multi-sleep modes



Figure 4: Different multi-sleep modes

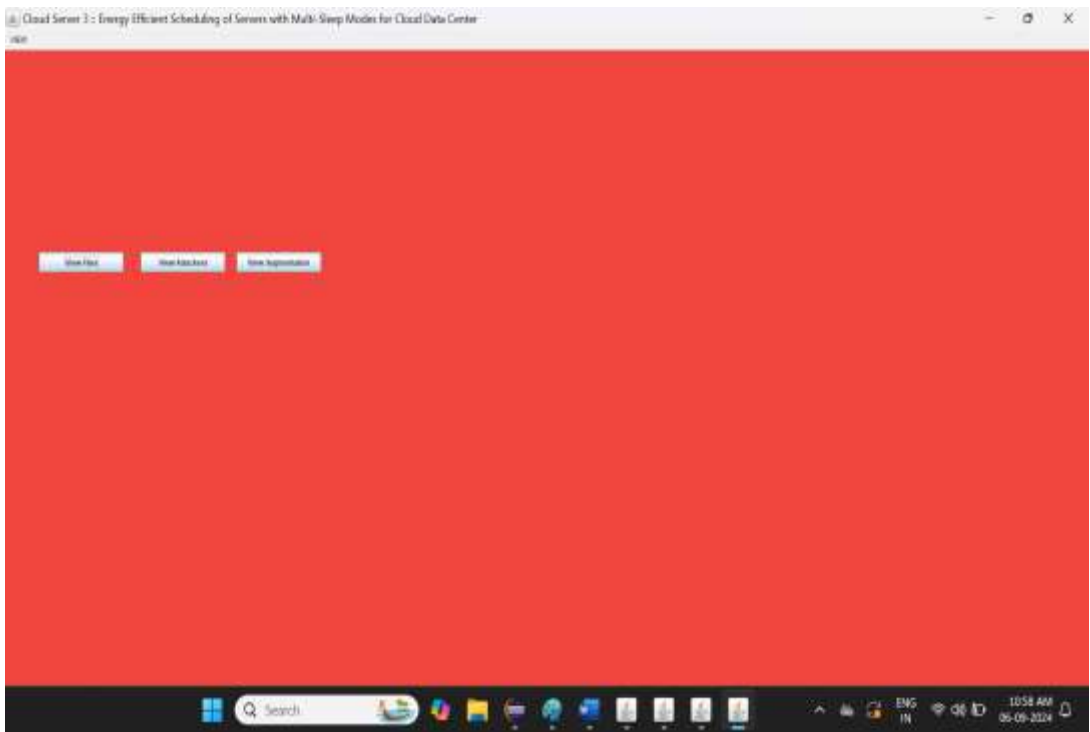


Figure 5: Different multi-sleep modes

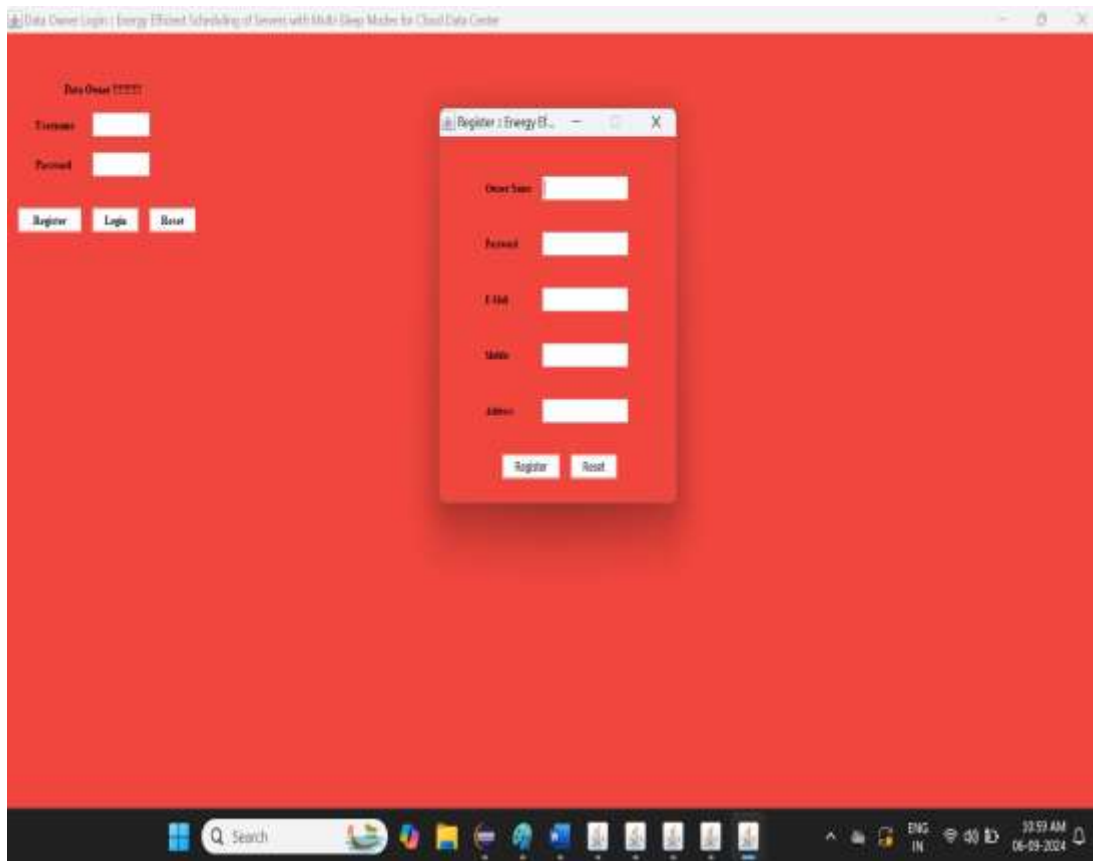


Figure 6: Register form 1

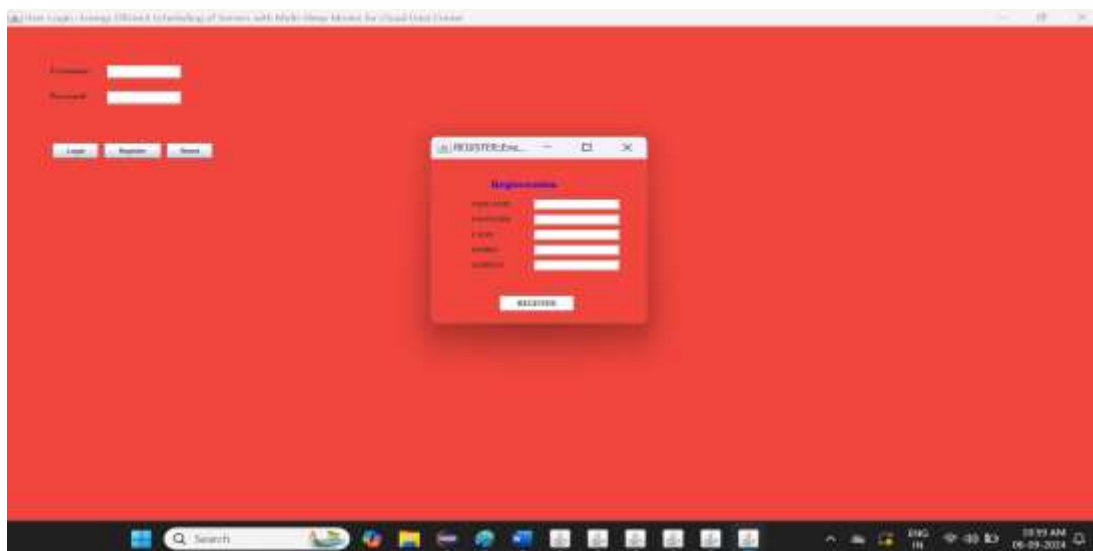


Figure 7: Register form 2

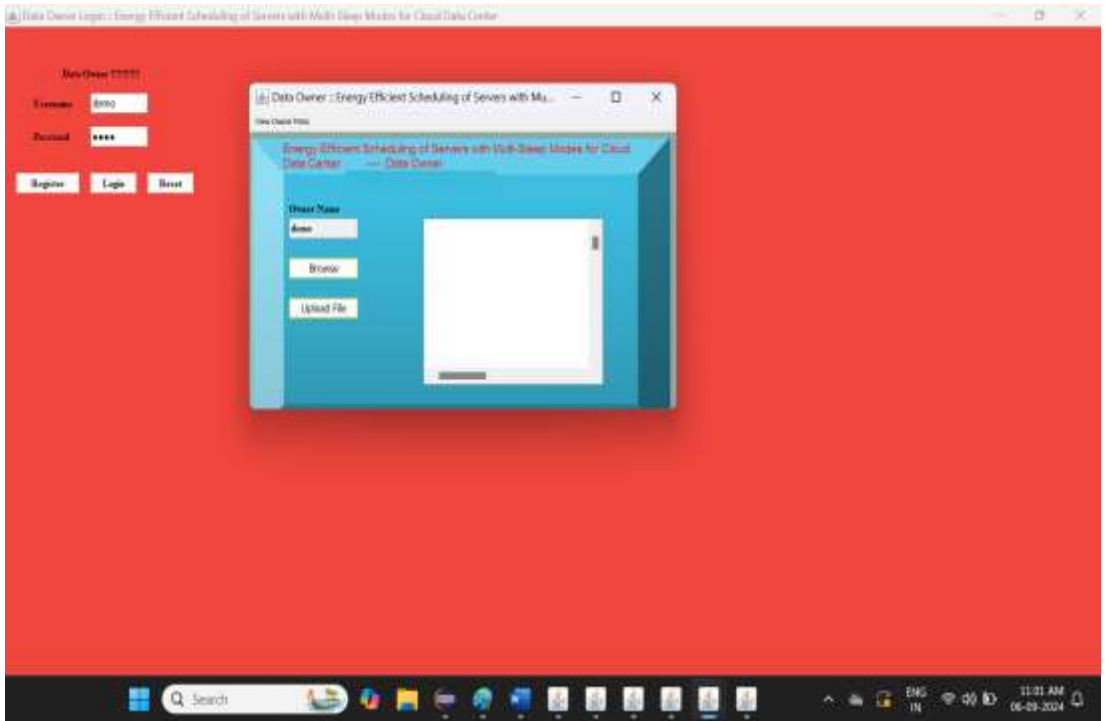


Figure 8: Energy Efficient of Data center

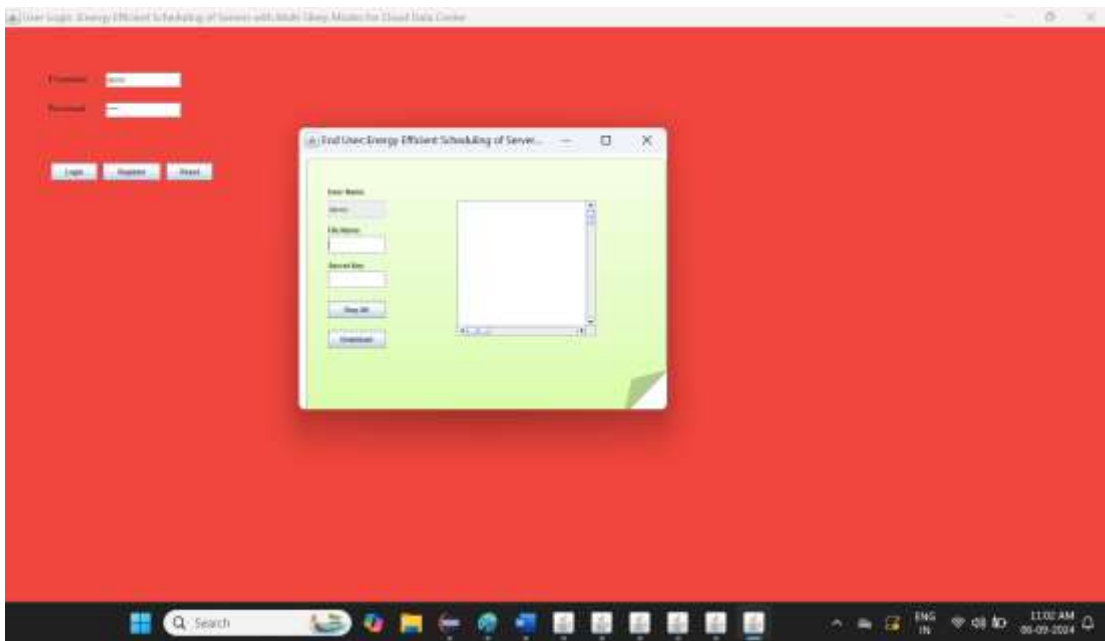


Figure 9: End user

VI CONCLUSION

In this paper, we studied the problem of scheduling of servers with multi-sleep modes for cloud data centers. The servers can make transitions between one active state and different sleep states, which involves different sleep power and transition delays for the sleep modes. We proposed Backtrack-and-Update method to make schedule of the servers, deciding how many servers in each state should be switched to which states in each timeslot, so that the total power consumption can be minimized while satisfying the QoS requirement. The problem is too large to be solved by existing methods, so we divide the whole problem and then conquer them one by one while considering the ongoing transitions during the breakpoints. We also consider using DVFS to further reduce the energy caused by the overprovisioned computing capacity. Experiments show that our scheduling using multi-sleep modes can significantly reduce the total energy with QoS of less than 10ms. Against the over-provisioned strategy of AlwaysOn, our method can reduce more than 28% of the total energy on average.

VII REFERENCES

- [1] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being green," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 211–222, Aug. 2012.
- [2] A. Gandhi, M. Harchol-Balter, and M. A. Kozuch, "Are sleep states effective in data centers," in *Proc. IEEE IGCC*, Jun. 2012, pp. 1–10.
- [3] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing electricity cost: Optimization of distributed internet data centers in a multielectricity- market environment," in *Proc. IEEE INFOCOM*, 2010, pp. 1–9.
- [4] Y. Zhang, Y. Wang, and X. Wang, "Greenware: Greening cloudscale data centers to maximize the use of renewable energy," in *USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing*, Dec. 2011, pp. 143–164.
- [5] S. Wang, Z. Qian, J. Yuan, and I. You, "A DVFS based energyefficient tasks scheduling in a data center," *IEEE Access*, vol. 5, pp. 13 090–13 102, 2017.
- [6] C. Gu, H. Huang, and X. Jia, "Green scheduling for cloud data centers using ESDs to store renewable energy," in *Proc. IEEE ICC*, Apr. 2016, pp. 1–6.
- [7] IBM, "CPLEX Users Manual," <https://www.ibm.com/support/knowledgecenter/SSSA5P12.7.0/ilog.odms.studio.help/pdf/uscplex.pdf>, 2017, [Online; accessed 25-December-2017].
- [8] M. C. P. T. L. T. C. Hewlett-Packard Corporation, Intel Corporation, "Energy Management of ACPI," <https://www.intel.com/content/dam/www/public/us/en/documents/articles/acpi-config-power-interface-spec.pdf>, 2017, [Online; accessed 25-December-2017].
- [9] D. G. Feitelson, D. Tsafir, and D. Krakov, "Experience with using the parallel workloads archive," *Journal of Parallel and Distributed Computing*, vol. 74, no. 10, pp. 2967–2982, Oct. 2014.
- [10] "Logs of real parallel workloads from production systems," <http://www.cs.huji.ac.il/labs/parallel>.
- [11] D. D. Gutierrez, "The Intelligent Use of Big Data on an Industrial Scale," <https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/>, 2017, [Online; accessed 25-December-2017].
- [12] J. Li, Z. Li, K. Ren, and X. Liu, "Towards optimal electric demand management for internet data centers," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 183–192, 2012.
- [13] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2, 2007, pp. 13–23.
- [14] Google, "Environmental Report," <https://static.googleusercontent.com/media/environment.google/zh-CN/pdf/google-2017-environmental-report.pdf>, 2017, [Online; accessed 25-December-2017].

- [15] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M. A. Kozuch, "Autoscale: Dynamic, robust capacity management for multi-tier data centers," *ACM Transactions on Computer Systems (TOCS)*, vol. 30, no. 4, p. 14, 2012.
- [16] D. Wong, "Peak efficiency aware scheduling for highly energy proportional servers," in *Proc. IEEE ISCA*, Jun. 2016.
- [17] A. Gandhi, M. Harchol-Balter, R. Das, and C. Lefurgy, "Optimal power allocation in server farms," in *Proc. ACM SIGMETRICS*, Jun. 2009, pp. 157–168.
- [18] M. E. T. Gerards, J. L. Hurink, and J. Kuper, "On the interplay between global DVFS and scheduling tasks with precedence constraints," *IEEE Trans. Computers*, vol. 64, no. 6, pp. 1742–1754, June 2015.
- [19] E. M. Elnozahy, M. Kistler, and R. Rajamony, "Energy-efficient server clusters," in *Proc. Springer PACS*, Feb. 2002, pp. 179–197.
- [20] M. Elnozahy, M. Kistler, and R. Rajamony, "Energy conservation policies for web servers," in *Proc. USENIX USITS*, Mar. 2003, pp. 8–8.