International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Energy Efficient Scheduling of Servers with Multi-Sleep Modes for Cloud Data Center

N Murali Krishna [1],Bitla Radhika[2], A Vineya[3],Celiveri Sai Kiran Goud[4], Vemula Nagasai Kashyap[5]
[2,3,4,5] UG Scholars, Department of CSE, **AVN Institute of Engineering and Technology**,Hyderabad, Telangana, India.
[1]Professor, Department of CSE, **AVN Institute of Engineering and Technology**, Hyderabad, Telangana, India.

ABSTRACT

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 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.

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. 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. 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

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

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) subproblem. 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.

LITERATURAL SURVEY

Are sleep states effective in data centers

While sleep states have existed for mobile devices and workstations for some time, these

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

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 na¨ıve 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., CO_2 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

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

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.

SYSTEM ANALYSIS

Existing System:

- ❖ 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 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 try to minimize energy cost through global DVFS on multi-core processors platform while considering the precedence constraint in task scheduling.
- ❖ Elnozahy 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. Rossi build power models to estimate the energy consumption of user applications under different DVFS policies.
- ❖ Florence first study the flow pattern of tasks of cloud, and then try to tune the incoming VM tasks with required frequency using DVFS. Lin use DVFS to reduce the power consumption in task scheduling in mobile cloud computing environment, but with no consideration of On/Off servers. Chen combine the three approaches of request dispatching, service management and DVFS to improve energy efficiency for large scale computing platform.

Proposed System:

- ❖ 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:

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

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 times lot.

IMPLEMENTATION

MODULES

Transition Delays and Power Consumption

ACPI (Advanced Configuration and Power Interface) has been an open industry standard, based on which modern servers are always designed with a power management hardware module, and it usually supports 5 sleep states from S1 to S5. In general, the transitions are only happening between sleep states and the active state S0. That is, the active state can be switched to any sleep state and a sleep state to the active state, but a sleep state cannot be switched to another sleep state. Note that, the transition delays and power consumption for different sleep modes are different.

QoS Model

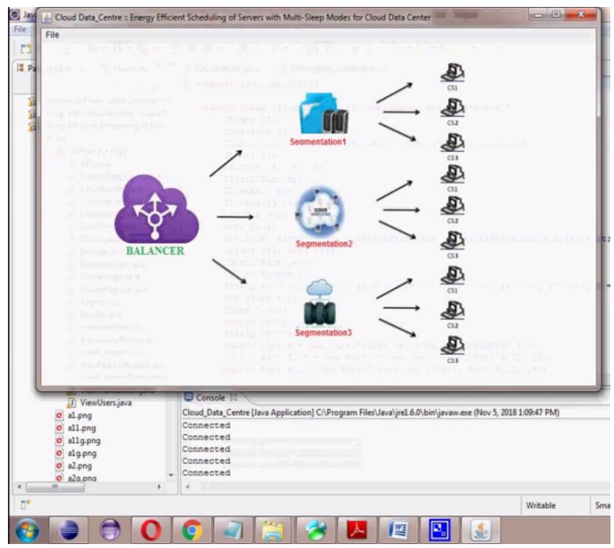
The quality of service (QoS) should always be considered in requests dispatching. In this part, we use M/M/n in queuing theory to model the response time of the incoming requests for a cloud data center. Let _t denote the number of incoming requests in t.

Total Power Consumption of Servers

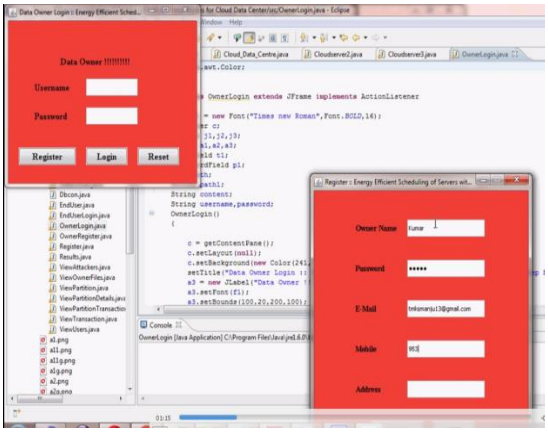
The power consumption of a server is closely related to its idle power and the current frequency level. Based on literature, the total server power using DVFS can be represented as follows:

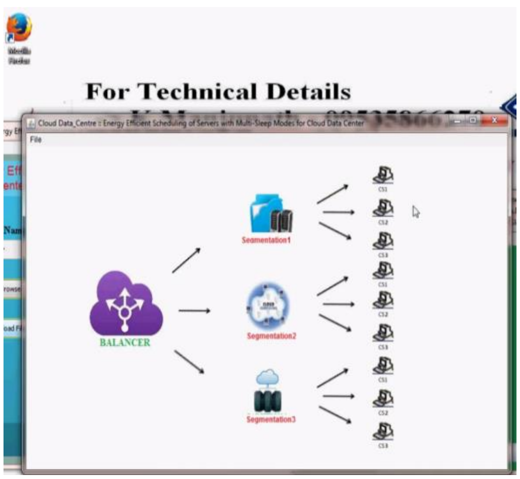
Pserver = Pidle + Pf (l); where Pidle denotes the idle power, Pf (l) denotes the dynamic power at frequency level l, respectively.

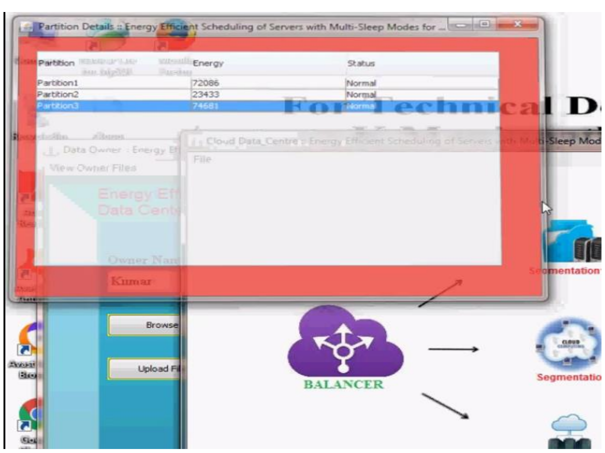
Results



International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362







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 over provisioned 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 Always On, our method can reduce more than 28% of the total energy on average.

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/SSSA5P 12.7.0/ilog.odms.studio.help/pdf/ usrcplex.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. Tsafrir, 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.