Energy Constrained Resource Scheduling for Cloud Environment

[1]R.Selvi , [2]S.Russia, [3]V.K.Anitha
[1]2nd Year M.E.(Software Engineering), [2]Assistant Professor
Department of IT
KSR Institute for Engineering and Technology, Tiruchengode, Tamilnadu, India

***Abstract*—Cloud Computing is used to access computing resources owned and operated by a third-party provider. It is internet-based computing to share resources, software and information. The rapid growth in demand for computational power has led to a shift to the cloud computing model established by large-scale virtualized data centers. Such data centers consume enormous amounts of electrical energy. However, to support green computing, cloud providers also need to minimize the cloud infrastructure energy consumption while conducting the service delivery. In existing system, dynamic server consolidation through live migration is an efficient way towards energy conservation in cloud data centers. It keeps the number of power-on systems as low as possible which is achieved by implementing PABFD algorithm with the help of MMT. The main drawback is that when the systems are in power off state will also increases power consumption when starts immediately. In proposed system, Power management strategies have been proposed for enterprise servers based on Dynamic Voltage and Frequency Scaling (DVFS). DVFS allows the server to transition the processor from high-power states to low-power states. The processors are assigned to deep sleep to reduce energy consumption. In deep sleep the server can be configured to use Direct Memory Access (DMA) to place incoming packets into memory buffers for processing in the active state. Request grouping will group received requests into batches and put the processor into sleep between the batches. The Virtual grouping scheme is enhanced to manage resources with load balancing mechanism. The system is improved with optimization mechanism to manage relative response time. Resource levels and application requirements are integrated in the allocation process. The system is adopted to support Dynamic Random Access Memory (DRAM) and Dual in-line Memory Module (DIMM) components.**

***Index terms*—Energy Management, Virtual Batching, Server Consolidation, DVFS, Power Management, Request Batching, Servers, Data Center**

I. INTRODUCTION

Data Centers have emerged as a back-bone infrastructure, housing large number of IT equipments such as servers, data storage, network devices, power and cooling devices etc. that facilitate the development of wide variety of services offered by the cloud. Currently, several service providers such as Amazon, Google, Yahoo, Microsoft, IBM and Sun, have their own data centers to provide the scalable services to a large customer base. Rapid development of IT industry and increasing demand for cloud services, the number of data centers has increased. These data centers consume enormous amount of energy to process its services resulting in increased energy consumption. The surging energy consumption of these data centers has become a serious concern from both economic and environmental standpoints.

According to McKinsey report, the energy consumption of data centers is $11.5 billion in 2010 and it doubles every five years. Gartner estimated that world wide IT infrastructures are responsible for 2% of global CO2 emissions and energy related costs account for the 12% of the total economical expenditures. The excessive energy consumption at data centers leads to high operational cost, large amount of CO2 emission and falling lifetime of hardware equipments. Hence, it is necessary to design energy-efficient data centers not only for ensuring system reliability but also reducing environmental impact and operational cost. Energy management techniques at the data centers can be static or dynamic. The static energy management techniques fail to address the run time adaptation of data centers in response to workload changes. The dynamic energy management techniques configure the data center at both hardware and software levels dynamically based on workload variability.

Further, the energy conservation can be achieved by efficient utilization of data center resources. Virtualization technology is one such powerful technology to address this energy inefficiency by increasing resource utilization. This technology allows multiple virtual machines (VMs) to share the resources on a single physical machine (PM). The features such as VM isolation and VM migration along with dynamic resource provisioning can be used either to consolidate virtual machines on fewer physical servers or to balance the load across physical servers in data centers, thereby ensuring applications' performance.

II. RELATED WORK

In data centers, if the physical machines consume more energy, these resources will also emit heat and harmful gases. This problem can be minimized by Green Cloud computing which provides the eco-friendly environment. The Green cloud computing reduces the energy consumption and also save energy.

A. Virtualization and Cooling technique

Authors in [1] provide another solution for greening the data centers. First is cooling system which minimizes the energy consumption. Various companies use river water for data center's cooling, open air data centers, air conditioned system etc. This system is expensive and not efficient for minimizing energy consumption. Second is virtualization technique in which more than one virtual machine loaded on a single physical machine. The virtualization technique provides the abstraction because the internal working hidden from the user. The user only accesses the web services and does not aware about the virtual machines. The virtualization techniques realize that a single physical machine is provided to the single user. This reduces the energy consumption because of single physical machine running. But performance will be degraded, the reason behind the performance degradation is a unbalanced load. Third is nanodata center technique which specifies that the large number of small sized data centers should be geographically distributed. Traditional data centers are of large size and few data centers are distributed and this technique consumes more energy because of long distance of data transmission. This nanodata center technique reduces the energy consumption because of the short distance between client and data center. These all techniques are used for greening the data centers.

B. Energy management in public and private cloud

In cloud computing there are two basic clouds first is public cloud and second is private cloud. Public cloud is accessible from any user through internet but private cloud is only accessible by the particular organization. The analysis of energy consumption is performed on basic web services such as storage as a service, software as a service and processing as a service [2]. Storage as a service provides a service in which user can store their data on cloud not on their personal machine. There is no need to buy any storage device such as hard disk, but the user have to pay according to the usage of the storage devices on cloud. Software as a service provides the latest software to the user through cloud for developing their own applications easily. There is no need to get license for software. Processing as a service is used for performing the computations on user's data and after all operations the result is provided to the user.

There are various energy consumption models which consumes energy. First, user equipments such as processor, memory, display unit etc. these devices consumes energy but at user side. Second, data center consumes energy because there are number of devices used for providing the services to the users. The energy consumption can be reduced by consolidating the servers but for consolidation the servers which are idle and have no task to perform can be turned off. In this method the load is distributed to the few servers and performance can be degraded. This process requires more attention.

The energy consumption analysis is performed on three web services such as storage as a service, software as a service and processing as a service. In case of storage as a service, the user creates their file and store on cloud. After some time if user wants to edit this file then the user must download the file form cloud and after providing the modifications again upload the file on cloud. This process consumes more energy because of uploading and downloading the files on cloud.

Figure 1 shows the comparison of public cloud and private cloud in which the public cloud consumes more energy than private cloud because of load on cloud. The private cloud is accessed by only the members of the organization but the public cloud is accessed by any user. In case of software as a service it also consumes energy for transporting the framework on user's machine through terminal. Last service is processing as a service consumes more energy in public cloud than a private cloud.

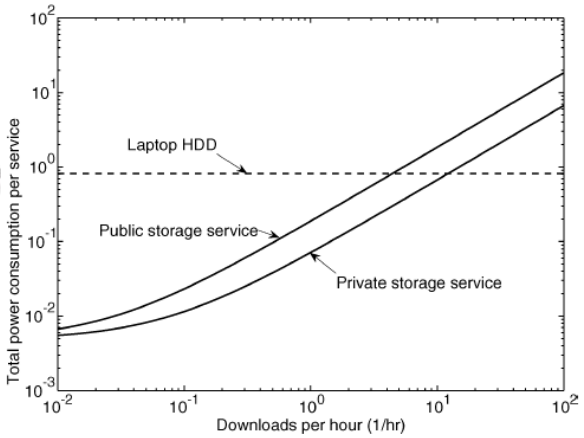

Fig.1 Comparison of Public and private cloud

C. Job Scheduling

In [3], scheduler schedules the tasks by determining the temperature of the task and node. The tasks are generated by Task Generation System. This system determines the temperature of the task by specifying the parameters such as initial temperature of the task, per minute rise in temperature and execution time of the task. This specification is given manually. After determining the temperature of the task then the prediction method is used for determining the temperature of the node. This prediction method uses two parameters: 1) task specification and 2) energy consumption. In this scheduler FCFS (First Come First Serve) algorithm and priority algorithm is used for scheduling. The priority algorithm schedules tasks according to the temperature of the task and node. The task, which requires low temperature, has high priority and the high temperature task have low priority. In this algorithm, one additional parameter is used for comparison which is a critical temperature and if any task requires temperature up to critical temperature then this task will not be executed, otherwise system gets failure. Figure 2 shows comparison chart with scheduler and without scheduler. Author says that system performance should not be impacted while energy consumption is being

minimized. Power aware virtual machine scheduling is another technique for reduction of energy consumption [4]. The virtual machines are scheduled according the power consumed by the virtual machines. This scheduling is provided for minimizing the performance overheads but with energy efficiency. But this technique does not providing the greenest data center which is the main aim of green cloud computing.

The job grouping is another technique for efficient energy consumption [5]. Jobs are scheduled according to the resource capability. Before the scheduling process, calculate the capability of each resource by selecting them. After calculating the capability resources then allocate the jobs to the resources according their capability. This scheduling technique is basically used for load balancing but with minimum reduction of energy consumption. In [6], the jobs are grouped together on the basis of similar resource requirement. This scheduling technique concentrates only on efficient resource management but with minimal reduction of energy consumption. This reduction is provided by reducing the waiting energy of the jobs.

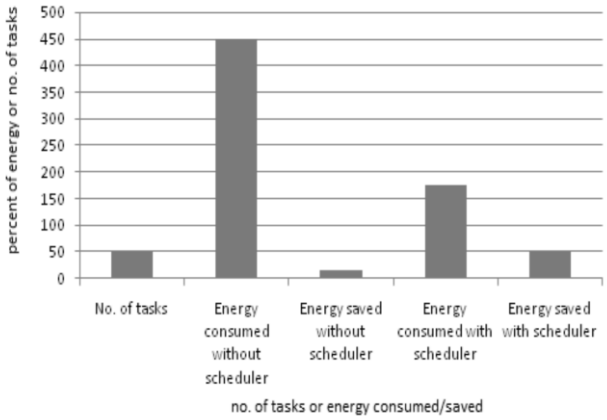

Fig.2 Saved power and energy consumed with/without scheduler

D. Resource Scheduling in Clouds

The need to provide a guaranteed level of service performance is important for data centers. This is largely due to a business model driven by strict service level agreements (SLAs) based on metrics such as response time, throughput, and reserve capacity. However, energy demands and associated costs are increasing at an alarming rate; it is projected that data centers in the US alone will consume 100 billion kWh of energy at a cost of 7.4 billion dollars per year by 2011. This poses a dilemma for data center operators; they must satisfy new and existing service contracts while minimizing energy consumption to reduce and strain on power generation facilities.

Data centers generally provision based on a worst-case scenario, which leads to a low-average server utilization in modern data centers. For example, a recent estimation suggests that the utilizations of web servers are often in the 5 to 12 percent range. These underutilized servers spend a large portion of their time in an idle state [5]. Several recent studies have shown that a server uses approximately 60 percent of its required peak power when it is idle. This over provisioning leads to large amounts of energy waste. Therefore, reducing energy waste, while guaranteeing SLA agreements, can lead to significantly reduced operating costs.

A well-known approach to addressing this problem is to transition the processor from high power states to low power states using Dynamic Voltage and Frequency Scaling (DVFS) whenever the performance allows. This approach effectively reduces the power consumption of the computer systems when the server has a medium intensity workload. However, the capability of this approach to reduce power consumption is limited when the server has a low-intensity workload due to two reasons. First, when the utilization of the processor is very low, the leakage power, which cannot be significantly reduced by DVFS, contributes a major portion of the power consumption. Second, many high performance processors only allow a small range of DVFS levels and even the lowest level provides a higher speed than is required for some light workloads. For example, in a case study on testbed, the power consumption of an idle server with an Intel Xeon 5360 processor can only be reduced from 163 to 158 W when the processor is transitioned from the highest DVFS level to the lowest one.

To further reduce energy consumption, processors need to be put into sleep states such as Deep Sleep. In Deep Sleep, the processor is paused and consumes significantly less power. For example, the power consumption of a server with an Intel Xeon 5500 Processor may be reduced to 23 percent of its peak value when the processor is switched to the Deep Sleep state [4]. When the processor is in Deep Sleep, the server can be configured to use Direct Memory Access (DMA) to place incoming packets into memory buffers for processing when the processor is returned to the active state, thus, avoiding harming the functionality of the hosted server applications. Therefore, to save more power for servers with light workloads, it can perform request batching to put the processor into the Deep Sleep state when there are few incoming requests. During the sleep time, delay and batch the requests when they arrive and wake the processor up when the earliest request in the batch has been kept pending for a certain batching time-out.

However, it is challenging to perform request batching directly on a virtualized server. Virtualization technologies such as Xen, VMware, and Microsoft Hyper-V allow provisioning multiple virtual machines (VMs) onto a single physical server. However, all the VMs on a single physical server are correlated due to sharing the same physical hardware, i.e., any state changes in the hardware affect all the VMs. Since different VMs may have different workloads and performance requirements, putting the processor into Deep Sleep based on the performance of one VM may affect the application performance of other VMs.

In this paper, proposed a Virtual Batching, a novel request batching solution for virtualized enterprise servers with primarily light workloads. This solution dynamically allocates the CPU resource such that all the VMs can have approximately the same performance level relative to their allowed peak values. Based on the uniform level, our solution then determines the time length for periodically batching incoming requests and putting the processor into sleep. When the workload intensity changes from light to medium, request batching is automatically switched to DVFS to increase processor frequency for performance guarantees.

Virtual Batching is also extended to integrate with server consolidation to achieve maximized energy conservation with performance guarantees for virtualized data centers. Server consolidation can improve server utilization by consolidating VMs onto a smaller number of servers on a long time scale. However, due to conservative resource profiling and various real-world constraints, servers after consolidation can still be underutilized. Virtual Batching can then be adopted to put the processors of active servers into sleep on a shorter time scale for further energy savings due to its much smaller overhead.

III. EXISTING SYSTEM

In the existing system, Local Regression method is used which sets the upper and lower threshold value for detecting the overloaded and underloaded host. The hosts which are above the threshold value are determined to be overloaded host also known as over-utilized host. The hosts which are below the threshold value are determined to be underloaded host also known as under-utilized host. Once it has been decided that a host is overloaded, the next step is to select particular VMs to migrate from this host.

To select which VMs to migrate from an under loaded/overloaded host, Minimum Migration Time (MMT) policy is applied. For VM to be placed on destination host is determined with the help of Utilization and Minimum Correlation(UMC) method which calculates the utilization of host and minimum correlation between the host. It is based on higher the correlation between applications that use the same resources on an over-utilized host, the higher probability the server to become overloaded.

Power aware load balancing strategy based on adaptive migration of VMs will be applied to virtual machines on cloud, considering higher and lower thresholds for migration of VM's on the servers and also RAM & Bandwidth for better performance & load balancing. If the load is greater or lower then defined upper & lower thresholds, VMs will be migrated ,boosting resource utilization of the cloud data center and reducing their energy consumption.

To reduce number of migration, MMT policy which is capable of reducing the number of migration and the energy consumption of virtual machine migration achieves load balancing and meet SLA requirements is used. To estimate VMs CPU utilization correlation, multiple correlation coefficients is used.

The UMC algorithm has two steps: (1)list *L* will be created from all hosts whose utilization is above threshold and by placing the VM on such host it will not become overloaded. Based on hill optimization method, the value of threshold is set. If list *L* is empty then the PABFD is applied, otherwise a host as candidate from this list will be selected. Hence it solves the overutilization/underutilization of datacenter host and also reduces energy consumption in data centers.

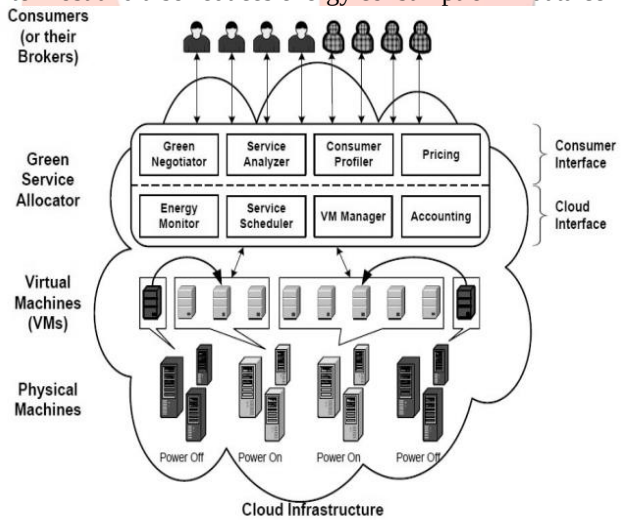

Fig.3 Architecture of Existing system

IV. *PROPOSED SYSTEM*

Request batching can be conducted to group received requests into batches and put the processor into sleep between the batches. Virtual Batching is a request batching solution for virtualized servers with primarily light workloads. The system dynamically allocates CPU resources with same performance level and peak values. Server consolidation is performed to fully utilize a small number of active servers in the data center. Static and dynamic server consolidation algorithms are used to assign data centers to the request batches. Static server consolidation algorithm is used for the offline mode in data centers. Online workload variations are managed by the dynamic server consolidation algorithms. Virtual batching is integrated with pMapper (power-aware application placement framework) to assign data centers for the workloads. The following drawbacks are identified in the existing system.

- Complex virtual server sleep process
- Average relative response time is not optimized

- Energy management is tuned for the processor
- Data center load is not managed

Virtual Batching with Memory Management

Cloud resources are provided for the users to perform computational tasks. Resources are provided by the providers with different servers. The servers are continuously running to provide resources. Energy consumption level for the servers is increased. Dynamic Voltage Frequency Scaling (DVFS) mechanism is used to adjust the voltage supply for the processors.

Power supply is managed under two states. They are sleep state and active state. The power supply is reduced in the sleep state. The resource requests are consolidated with request levels. In the same way the server resource levels are also consolidated with availability details. The request and server consolidation mechanism is used in the Virtual batching technique. The Virtual batching technique is used to manage energy levels in cloud resource sharing environment. The Virtual batching mechanism is enhanced to manage memory resources. The system is designed with the following objectives.

- To manage cloud resources with energy consumption levels
- To apply Virtual Batching technique for energy constrained resource managed process
- To group relevant requests and server resources
- To enhance Virtual Batching with load balancing mechanism
- To incorporate workload variations in dynamic server consolidation process
- To provide power management on memory devices
- To increase the average relative response rate

The Virtual Batching scheme is enhanced to manage resources with load balancing mechanism. The system is improved with optimization mechanism to manage relative response time. Resource levels and application requirements are integrated in the allocation process. The system is adopted to support Dynamic Random Access Memory (DRAM) and Dual in-line Memory Module (DIMM) components.

The Virtual Batching scheme is improved to manage power for computational and storage units. Request consolidation is improved with optimization techniques. Request load is distributed with different servers. The system is divided into five major modules. They are resource management, consolidation process, resource allocation process, load balancing process and power management on memory units. Resource management module is designed to maintain the resource availability under the providers. Request and server consolidation tasks are carried out under the consolidation module. Resource allocation module handles the scheduling process. Request loads are distributed under load balancing module. Memory devices are managed with power usage levels with memory management module.

V. CONCLUSION

Cloud resources are managed with energy consumption levels. Virtual Batching scheme is used to allocate resources with request and server consolidation. The system is improved with optimization schemes to increase response rate. The system is enhanced to support energy management under memory devices. High server utilization is achieved in the system. Energy consumption is minimized by the resource scheduling scheme. The system achieves efficient application performance. The system maximizes the throughput in resource sharing process.

REFERENCES

[1] D. Cavdar, F. Alagoz, "A Survaey of Research on Greening Data Centers", IEEE Symposium on Selected Areas in Communications, pp. 3237-3242, 2012.

[2] J. Baliga, R. Ayre, K. Hinton, R. Tucker, "Green Cloud Computing: Balancing Energy in Processing, Storage and Transport", Proceedings of the IEEE, Vol. 99, pp. 149-167, 2011.

[3] S. Arora, V. Chopra, "A Predictive Energy Aware Hybrid Resource Scheduler for Green Cloud Computing", International Journal of Applied Research in Computing, Vol. 1, pp. 1-5, 2013.

[4] Y. Ponnusamy, S. Sasikumar, "Application of Green Cloud Computing for Energy Management", International Joural of Computer Science &
Research in Computing, Vol. 1, pp. 1-5, 2013.

[5] S. Selvarani, "Improved Job-Grouping based PSO algorithm for Task scheduling", International Journal of Engineering Science and Technology, Vol. 2(9), pp. 4687-4695, 2010.

[6] R. Nanduri, N. Maheshwari, R. Raja, V. Varma, "Job Aware Scheduling Algorithm for MapReduce Framework", IEEE International Conference on Cloud Computing Technology and Science, pp. 3-4, 2011.