A Survey Paper on Energy Consumption in Green Cloud Computing by DENS Algorithm

Raj Kiran. K\textsuperscript{1} & Hamsaveni. M\textsuperscript{2}

\textsuperscript{1}M.Tech Student, \textsuperscript{2}Assistant Professor,

\textsuperscript{1,2} Vidyavardhaka College of Engineering, Mysuru, Karnataka.

Abstract: Cloud is a managed pool of resources which provide on-demand services or resources to the remote users over a network, evenly can increase the bandwidth as per necessary (elasticity) and pay as per usage. Datacentres consume large amounts of energy for computation of data, data processing, storage and communication, which directly impacts the environment through heat, carbon emissions and many other factors [1]. As a result, cloud service providers are seeking innovative ways that allow them to reduce the energy consumption. In this survey paper, which states some algorithm aimed at reducing energy consumption in cloud computing datacentres, likely ESWCT (Energy-aware Scheduling algorithm using Workload-aware Consolidation Technique) [7], DVFS (Dynamic Voltage Frequency Scaling) [4,7,9], DENS (data centre energy-efficient network-aware scheduling) [11]. The results showed a significant improvement in reducing the energy and improvement in resource utilisation.

Keywords—Cloud Computing; Green Computing; power management; virtualization; Energy efficiency; Interconnected Fiber Channel; Switches

1. Introduction

In order to provide cloud services to all the end users, this impacts in increasing amount of energy consumption for data computation, processing, communication and storage of data. This leads to excess usage of power and heat in data centres. A survey says that "in the USA power consumption is about 124 billion kilowatt-hour as per 2016 and this will be around 170 billion kilowatt-hours, as of worldwide in 2020 [32-33], which produces 170 million metric tons of carbon emission per year."

[5,22,34] Therefore emerging ideas to minimizing the energy consumptions where many algorithms are proposed namely DVFS, ESWCT, DENS and many others. Energy-aware Scheduling Algorithm using Workload-aware Consolidation Technique (ESWCT) [7] centralized algorithm focuses on consolidating the VMs in a minimum amount of servers, based on balancing the integrated resource utilisation means processor, memory and bandwidth of the network, as these are the concerns which are concurrently shared by the users in datacentres.

Dynamic Voltage Frequency Scaling (DVFS) [4,7,9], which is applied by number of special processors which has the capability to operate at different voltage levels and frequency levels. An appropriate supply voltage and frequencies of processing elements are selected in order to minimize energy consumption, which doesn't violate Service Level Agreement (SLA). Datacentre Energy-efficient Network-aware Scheduling (DENS) [11] algorithms are to aim at reducing the energy consumption in a datacentre by optimizing the trade-off between the task consolidation and traffic pattern distribution, based on multiple weighted computational functions.

2. Literature Survey

Generally, in green cloud computing, a central scheduler is responsible for allocation and migration of (Virtual Machine) VMs. During peak hours, the centralized entity will be overloaded or very heavily loaded, in some cases it can be halted [1,6,12-18]. Somehow this entity is responsible for the phase of the server, either in allocation phase or migration phase.

In our DENS algorithm [11], each server is responsible for admitting and migrating VMs, and the servers are responsible for deciding its phase of operation that's either in allocation phase or consolidation/migration phase, by this distribution entity over servers improve resource, response time and maintaining loads.

2.1. Energy-aware Scheduling Algorithm using Workload-aware Consolidation Technique

The aim ESWCT algorithm is for managing cloud data centers to make full use of the resources. These are based on heterogeneous workloads need a variety of resources in cloud data centers simultaneously. These algorithms [11] results in higher utilisation of resources and reduces the power consumption. Framing the placement of VMs is optimized to minimize power consumption and maximize performance. The algorithms do not handle strict SLA requirements. SLA can be violated due to the variability of the workload. This algorithm
has a technique for minimization of power consumption, concentrating the workload to the minimum of physical nodes and switches by making the idle nodes off.

Algorithm periodically monitors a load of resources (CPU, disk storage, and network interface) and eventually makes decisions on switching nodes on/off to minimize the overall power consumption. The algorithm runs, on a master node, in some cases due to heavy load in larger systems, this may become a performance bottleneck, which creates a Single Point of Failure (SPF). The problem arises when it comes to providing strict SLAs ensuring no performance degradation, which is required for a Cloud data center.

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 [19]. The system maintains an active set of servers selected to serve requests for each service. The network switches are reconfigured dynamically to change the active set of servers whenever necessary (request occurs). Energy consumption is reduced by switching idle servers to power saving modes either sleeps or hibernates [1,10]. The latency during switching nodes on/off is not taken into account.

This is 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. Average utilisation of CPU, memory and network bandwidth during each observed period, is taken care respectively.

2.2. Dynamic Voltage Frequency Scaling

Dynamic Voltage and Frequency Scaling (DVFS) [2-3], this algorithm is based on a bin packing problem where servers are represented as bins with variable sizes due to the frequency scaling. In this algorithm, VMs are consolidated in those hosts that have a high utilisation but, on the contrary, have a low increase in frequency due to the utilisation increment. This approach minimizes the number of bins used and proposes two contributions:

1) A DVFS policy [2] that takes into account the trade-offs between energy consumption and performance degradation.

2) A novel consolidation [3] algorithm that is aware of the frequency that would be necessary when allocating a Cloud workload in order to maintain QoS.

Dynamic Voltage and Frequency Scaling (DVFS) helps to reduce the consumption of underutilized resources dynamically, the static consumption by reducing the number of active servers, thus increasing their utilisation. Cloud services are provided under strict Service Level Agreement (SLA) conditions, power consumption in data centers may be minimized, without violating the SLA requirements whenever it is feasible.

Whereas migration is to allocate in one particular virtual machine so as to reduce the wastage of energy and to provide within a particular time limit [19-21]. Somewhat, we integrate the concept called dynamic voltage frequency scaling in CPU utilisation model that specifies the frequency to proceed each task to complete within the particular deadline.

Three different types of services, energy saving mainly deal with the infrastructure as a service.

Etienne Le Sueur [35-37], this technique is commonly used in power management where the processor of the clock frequency is decreased to allow the reduction in voltage supply. This technique is widely used to reduce energy consumed in varying processors by the voltage and frequency at a run time the main purpose of reducing the energy efficiency and power during the various stages according to the workload and capacity.

Lin gong [27-30], this paper is about saving the energy strategy concept. It is to develop the application to reduce and calculate the different frequency in CPU host. Here the data center is to maintain the different virtual machine and to reduce power, such energy saving in the green computing framework reduces the energy consumed in order to increase the performance of the quality of service, reliability, availability of the servers. Here it also mentions the two terms for this green cloud computing concept is cloud computing and platform computing.

Cloud computing [28] is a model where it enables the on-demand process that shares computing resources in the application and the services.

Platform computing is based on the service-oriented program for both private and public cloud where it proposes the window live and window azure concept of the energy saving approach.

Ashima Agarwal [23-26], the concept of migration in a virtual machine is scheduled in particular process. In order to provide a progress to migrate, and it also creates a large amount of availability to reduce the balancing while we migrate [24] the data from one particular VM to other as process hence it also a benefit to control all the data to be saved and copied respectively.

Fig.1. the aim of this flowchart is to provide an energy conservation to save the data from one VM to the other VM using the green cloud technique. Whereas users submit the job length (deadline), here the term deadline represents the maximum length of the job in execution time [25] for which the algorithm checks the least frequency to calculate the time and job length. It majorly will depend on the job length to migrate either to one virtual machine or in several other virtual machines that consume comparatively lesser amount of energy.
3. The Survey Algorithm

3.1. Datacentre Energy-efficient Network-aware Scheduling Algorithm (DENS)

Other than DENS all the other algorithms are focusing only on job distribution between computing servers based on workload, power optimization or heating issues. This paper says the role of communication fabric in data center energy consumption and presents a scheduling approach that combines energy efficiency and network awareness, named DENS. The DENS [11] methodology balances the energy consumption of a data center, individual job performance, and traffic demands. The DENS approach optimizes the trade-off between job consolidation (to minimize a number of computing servers) and distribution of traffic patterns (to avoid hotspots in the data center network).

The slice of roughly 40% is related to the energy consumed by information technology (IT) equipment, which includes energy consumed by the computing servers as well as data center network hardware used for interconnection. In fact, about one-third of the total IT energy is consumed by communication links, switching, and aggregation elements, while the remaining two-thirds are allocated to computing servers. Early solutions implemented distributed algorithms for making data center hardware energy efficient.

There are two popular techniques for power savings in computing systems.

1. The Dynamic Voltage and Frequency Scaling (DVFS) technology, adjusts hardware power consumption according to the applied computing load.

2. Dynamic Power Management (DPM) achieves most of the energy savings by powering down devices at non-runtime. To make DPM scheme efficient, a scheduler must consolidate data center jobs on a minimum set of computing resources to maximize the amount of unloaded servers that can be powered down (or put to sleep). Because the average data center workload often stays around 30%, the portion of unloaded servers can be as high as 70%.

Two large traffic flows may be assigned to share the same path if their hash values collide leaving other paths under-loaded. The problem is solved with the introduction of a complex central scheduler that performs flow differentiation and analysis of flow traffic demands across the data center network. The approach presented in, also allows job migration control during runtime with a specifically designed network-aware scheduler. The migration scheduler is aware of the migration delays and bandwidth resources required. The network-awareness refers to the ability of DENS approach to receive and analyze a run-time feedback from the data center switches and links as well as take decisions and actions based on the network feedback. The DENS [11] methodology aims to achieve the balance between individual job performances, job QoS requirements also defined in Service Level Agreement (SLA), traffic demands, and energy consumed by the data center. Data intensive jobs [19] require the low computational load but produce heavy data streams directed out of the data center as well as to the neighboring nodes.

In this methodology, the network awareness is achieved with the introduction of feedback channels from the main network switches. Moreover, this methodology reduces computational and memory overhead compared to previous approaches, such as flow differentiation, which makes the DENS methodology easy to implement and port to existing data center schedulers.
could typically 100,000 hosts and the requirement to
keep layer-2 switches in the access network, a three-
tiered design becomes the most appropriate option.
Although 10 Gigabit Ethernet (GE) transceivers are
commercially available, in a three-tiered architecture
the computing servers (grouped into racks) are
interconnected using 1 GE links. This is due to the
fact those 10 GE transceivers:
(a) are too expensive and
(b) comparably offers more capacity than needed
for connecting computing servers.

In current data centers, rack connectivity is
achieved with inexpensive Top-of-Rack (ToR)
switches. A typical ToR switch shares two 10 GE
uplinks with 48 GE links that interconnect
computing servers within a rack. The difference
between the downlink and the uplink capacities is
equal to 48:20 = 2.4: 1. Therefore, under full load,
only 416 Mb/s will remain available to each of the
individual servers out of their 1 GE links. At the
higher layers of hierarchy, the racks are arranged in
modules with a pair of aggregation switches
servicing the module connectivity. The bandwidth
between the core and aggregation networks is
distributed using a multi-path routing technology,
such as the Equal Cost Multi-Path (ECMP) routing.
The ECMP technique performs a per-flow load
balancing, which differentiates the flows by
computing a hash function on the incoming packet
headers. For a three-tiered architecture, the
maximum number of allowable ECMP paths bounds
the total number of core switches to eight.
Nevertheless, we must note that all of the findings of
this research will remain valid for any or all types of
data center topologies.

- Energy models
Computing servers account for a major portion of
data center energy consumption. The power
consumption [22] of a computing server is
proportional to the CPU utilisation. An idle server
consumes around two-thirds of its peak-load
consumption to keep memory, disks, and I/O
resources running. The remaining one-third changes
almost linearly with the increase in the level of CPU
load. The DVFS scheme adjusts the CPU power
according to the offered load. The scope of the
DVFS optimization is limited to CPUs. Therefore,
computing server components, such as buses,
memory, and disks remain functioning at the original
operating frequency. On the other hand, the DPM
scheme can power down computing servers, which
makes such a technique very energy efficient.
However, if there is a need to power up (powered
down) the server, a considerable amount of energy
must be consumed compared to the DVFS scheme.
Switches form the basis of the interconnection fabric
that delivers job requests to the computing servers
for execution. Energy consumption of a switch
depends on the:
(a) Type of switch,
(b) Number of ports,
(c) Port transmission rates, and
(d) Employed cabling solutions.

Obviously, not all of the switches can dynamically
be put to sleep. Each core switch consumes a
considerable amount of energy to service large
switching capacity. So it is advisable to keep the core
network switches running all the time at their
maximum transmission rates. On the other hand, the
aggregation switches service modules, it makes
perfect sense to power down unused aggregation
switches. However, such an operation must be
performed carefully by considering possible
fluctuations in job arrival rates. Typically, it is
enough to keep a few computing servers running idle
on top of the necessary computing servers as a buffer
to account for possible data center load fluctuation.

- Data center tasks models
In cloud computing, incoming requests are
typically generated for such applications like web
browsing, instant messaging, or various content
delivery applications. The majority of such requests
can be classified according to the amount of
computing and communications they require into
three categories: – Computationally Intensive
Workloads (CIWs) model, Data-Intensive Workload
(DIWs) model and Balanced Workloads (BW).

CIWs demand a large amount of computing
resources but produce almost no data transfers in the
interconnection network of the data center. The
process of CIW energy efficient scheduling should
focus on the server power consumption footprint
trying to group the workloads at the minimum set of
servers as well as to route the traffic produced using
a minimum set of routers. There is no danger of
network congestion due to the low data transfer
requirements, and putting the most of the switches
into the sleep mode will ensure the lowest power of
the data center network.

Data-Intensive Workloads (DIWs) produce
almost no load at the computing servers but require
heavy data transfers. DIWs aim to model such
applications like video file sharing where each
simple user request turns into a video streaming
process. As a result, the interconnection network and
not the computing capacity becomes a bottleneck of
the data center for DIWs. Fortunately, there should
be a continuous feedback implemented between the
network elements (switches) and the central
workload scheduler. Based on such feedback, the
scheduler will distribute the workloads taking current
congestion levels of the communication links. It will
avoid sending workloads over congested links even
if certain server's computing capacity will allow
accommodating the workload. Such scheduling
policy will balance the traffic in the data center.
network and reduce average time required for a task delivery from the core switches to the computing servers.

Balanced Workloads (BW) [33] aim to model the applications having both computing and data transfer requirements. BWs load the computing servers and communication links proportionally. With this type of workloads, the average load on the servers equals to the average load of the data center network. BWs can model such applications as geographic information systems which require both large graphical data transfers and heavy processing. Scheduling of BWs should account for both servers' load and the load of the interconnection network.

- Data center network congestion

Utilizing a communication fabric in data centers entails the concept of running multiple types of traffic (LAN, SAN, or IPC) on a single Ethernet-based medium. On one side, the Ethernet technology is cheap, easy to deploy, and relatively simple to manage, on the other side, the Ethernet hardware is less powerful and provisions for small buffering capacity. Any of the data center switches may become congested either in the uplink direction or the downlink direction or both. In the downlink direction, the congestion occurs when individual ingress link capacities overcome individual egress link capacities. In the uplink direction, the mismatch in bandwidth is primarily due to the bandwidth oversubscription ratio, which occurs when the combined capacity of server ports overcomes a switch's aggregate uplink capacity. Congestion (or hotspots) may severely affect the ability of a data center network to transport data. The IEEE 802.1Qau specifications introduce a feedback loop between data center switches for signaling congestion. Such a feedback allows overloaded switches to hold off heavy senders from sending the congestion notification signal. Such a technique may avoid congestion-related losses and keep the data center network utilisation high. However, it does not address the root of the problem as it is much more efficient to assign data-intensive jobs to different computing servers in the way that jobs avoid sharing common communication paths. This trade-off between energy efficiency, data center network congestion, and performance of individual jobs is resolved using a unified scheduling metric presented in the subsequent section.

4. Algorithm

One of the algorithm says about the scheduler on/off when the switch is idle, thus checks the current load of the server, and accepts the load. If the load on the running server is more (threshold), then allocate a new server in the system of the datacenter. This algorithm based on the workload.

- It works on Least Loaded Server (LLS).
- It periodically checks total free resources and reports Dynamic Power Management to shut it down.
- Allocation Phase, just allocates by turning switches and servers on.
- Consolidation Phase, just checks what is the largest VM can fit in the server.

If there is no sufficient space for the system to work then, it goes to Migration Phase.

Algorithm-1:- DENS Algorithm.

Algorithm-1:-

The DENS methodology minimizes the total energy consumption of a data center by selecting the best-fit computing resources for job execution based on the load level and communication potential of data center components. The communicational potential is defined as the amount of end-to-end bandwidth provided to individual servers or group of servers by the data center architecture. Contrary to traditional scheduling solutions that model data centers as a homogeneous pool of computing servers, the DENS methodology develops a hierarchical model consistent with the state of the art data center topologies.

For a three-tier data center, we define DENS metric \( M \) as a weighted combination of server-level \( f_s \), rack-level \( f_r \), and module-level \( f_m \) functions:

\[
M = \alpha \cdot f_s + \beta \cdot f_r + \gamma \cdot f_m
\]

Where \( \alpha \), \( \beta \) and \( \gamma \) are weighted coefficients that define the impact of the corresponding components (servers, racks, and modules) on the metric behavior. Higher \( \alpha \) value favor the selection of highly loaded servers in light racks.

Higher \( \beta \) values will prioritize computationally loaded racks with low network traffic activity. Higher \( \gamma \) values favor selection of loaded modules. The \( \gamma \) parameter is an important design variable for job consolidation in data centers. This DENS is inspired by the Random Early Detection (RED) and
Backward Congestion Notification (BCN) technologies.

5. Performance and Statistical Analysis

For performance evaluation purposes, the DENS methodology was implemented in the Green Cloud simulator. Green Cloud is a cloud computing simulator developed by us to capture data center communication processes at the packet level. Green Cloud offers users a detailed fine-grained modeling of the energy consumed by the elements of a data center, such as servers, switches, and communication links.

Fig. 3. The Structure of Green Cloud Simulator.

For example, fig.3 for data rates of up to 1 GB/s energy profiles of network links and switches' transceivers are driven by twisted pair technology while for greater rates of 10 GB/s optical multimode transmitters are used. For workload execution, Green Cloud employs deadline based model, i.e. each task should be able to perform a specified amount of computations and transmit a given amount of data before a specified deadline for successful completion.

5.1. Some Statistical Values

Fig. 4. Average processor utilization per server in small datacentre.

Fig.4. presents the average processor utilisation result in a small datacentre, consisting of 400 servers. It shows that the DENS algorithm has significantly improved the processing utilisation when compared to ESWCT. It has outperformed DVFS processing utilisation with more than 18%. The figure shows that the benchmarks were not robust enough to schedule heavy loads so the system halts. DVFS was able to schedule up to 900 VMs only, whilst ESWCT allocated up to 1300 VMs. On the other hand, our DENS algorithm was more robust and successfully scheduled all loads in this experiment.

Fig. 5. Average processor utilization per server in large datacentre.

Fig.5. presents the average processor utilisation results in a large datacentre consisting of 1,000 servers. It shows the DENS algorithm as having significantly improved the processing utilisation more than ESWCT. Furthermore, it has improved the processing utilisation 17% more than DVFS.

Fig. 6. The total energy consumption in small datacentre.

Fig.6. shows the energy consumption in a small datacentre, consisting of 400 servers. It shows that the DENS algorithm has significantly reduced the power consumption compared with the benchmarks. In heavy loads, the benchmarks were not robust enough to allocate all loads, meaning the system halts. DVFS was able to allocate up to 900 VMs prior to halting, whilst the ESWCT halted at 1300 VMs. Our DENS algorithm was robust enough to allocate all loads in this experiment.
Fig. 7. The total energy consumption in small datacentre.

Fig. 7 shows the energy consumption through the DENS algorithm and the benchmark algorithms in a large datacentre of 1,000 servers. It shows that the DENS algorithm has significantly reduced the energy consumption better than DVFS considering all loads up to 60%. Furthermore, it reduced energy consumption better than ESWCT—up to approximately 70%.

6. Conclusion

DVFS was able to allocate up to 900 VMs prior to halting, whilst the ESWCT halted at 1300 VMs [8]. DENS algorithm was robust enough to allocate all loads in this experiment. This DENS algorithm explains the role of communication fabric in data center energy consumption and presents a methodology, termed DENS, that combines energy-efficient scheduling with network awareness. The DENS methodology balances the energy consumption of a data center, individual job performance, and traffic demands. The DENS approach optimizes the trade-off between job consolidation (to minimize a number of computing servers) and distribution of traffic patterns (to avoid hotspots in the data center network). DENS methodology is particularly relevant in data centers running data-intensive jobs which require the low computational load but produce heavy data streams directed to the end-users.

7. Future Enhancement

Future work will focus on the implementation and testing of DENS methodology in realistic setups using test beds. The design and specification of DENS metric are tight to the underlining data center architecture. In this paper, three-tier architecture is used along which that DENS and DPM are used. However, the adaptation of DENS approach to other existing and upcoming data center architectures is already on-going.

8. References


