ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 6, Issue 4, April 2017

A Brief Review of Energy Efficient Techniques in Data Centre's

Jagjeet Singh , Manju Ahuja

***Abstract*—**
The invention of Cloud Computing has rapidly changed the paradigm of ownership-based computing approach to subscription-oriented computing by providing access to scalable infrastructure and services on-demand. However, Clouds are essentially datacenters hosting application services offered on a subscription basis. They require high energy usage to maintain their operations. Today's data centers are primarily used for internet or network-based activities. They contain servers that store and process electronic data, communicate with other computer networks, and/or manage user interactions with server-based software tools and web portals. Achieving reliability through power/cooling redundancy and the use of UPS and ATS devices increase the electrical energy used by a data center. Many metrics have been introduced to develop an understanding and comparison of energy use and efficiency in data centers. Metrics focused to improve power utilization efficiency to reduce the overall costs for operations.

***Index Terms*— Server Energy Efficiency, Metrics, Power Usage Effectiveness, Data Center Architecture**

I. INTRODUCTION

Cloud is actually refers to network, often represented by a cloud network. Ones can be presented as a layered architecture that can be viewed as a collection of IT services referred to (SaaS) Software-as-a-Service, (PaaS) Platform-as-a-Service, and Infrastructure-as-a-Service (IaaS). Among all these SaaS permit users to run applications remotely from the cloud [1]. Cloud Computing provides a greatly scalable and cost-effective computing infrastructure for running IT applications such as High Performance Computing (HPC). The Cloud users can store, access, and share any amount of information online. likewise, small and medium enterprises/organizations do not have to worry about purchasing, configuring, administering, and maintaining their own computing infrastructure. They can in its place focus on improving their core competencies by exploiting a number of Cloud Computing benefits such as low cost, datacenter efficiencies, on-demand computing resources, faster and cheaper software development capabilities. According to a report published by the European Union, a decrease in emission volume of 15-30% is required before the year 2020 to keep the global temperature increase below 2oC. Thus, the rapidly growing energy consumption and CO2 emission of Cloud infrastructure has become a key environmental concern. Hence, energy proficient solutions are required to ensure the environmental sustainability of this new computing paradigm [2]. Data centers have developed into major energy hogs. Interestingly, one key aspect in the thermal management of a data center is still not very fine understood: controlling the set point temperature at which to run a data center's cooling system. Data centers usually operate in a temperature range between 20C and 22C, some are as cold as 13C degrees. While increasing data center temperatures might seem like an easy way to save energy and reduce carbon emissions, it comes with some concerns, the most apparent being its impact on system reliability [3].

A. Server Energy Efficiency

ATA centers often comprise thousands of enterprise servers that normally serve millions of users globally in a 24-7 fashion. The increasing demand for computing resources has recently facilitated the rapid proliferation and growth of data center facilities. Until lately, data centers have focused mainly on providing the desired performance. As a result, raw throughput increased tremendously. However, today's data centers consume a huge amount of electrical power. A strategy is used to reduce server energy consumption, in a way that is aware of the interactions among power, temperature, leakage, and workload dynamics. It includes:-
-Empirical models to estimate various power components in the server (e.g., static and dynamic power, CPU and memory power).
-It analyzes leakage vs. cooling power tradeoffs at the server level, and show the importance of temperature-dependent leakage in server energy consumption. It also study the relationship among power, temperature, application characteristics and workload allocation.
- A control strategy that dynamically sets the optimum cooling for arbitrary workloads. Compared to prior techniques this policy reduces leakage plus fan energy by an additional 3% without any impact on performance [4].

B. Metrics to develop an understanding of energy use and efficiency in data centers

Infrastructure efficiency metrics have to consider the energy consumption correlated with effective utilization. The metrics are measured performance at a point in time or averaged performance over the measurement time frame: Many metrics have been introduced to develop an understanding and comparison of energy use and competence in data centers. Two popular metrics are:
• IT productivity Per Embedded Watt (IT-PEW), which captures the power competence of the servers and is a metric that operators should want to maximize (Brill, 2007); and
• PUE, which is analogous to Data Center infrastructure Efficiency (DCiE) and Site Infrastructure Energy Efficiency (SI-EER) ; all three of which attempt to quantify the power efficiency of the infrastructure systems and are metrics operators should want to minimize.

C. Power Usage Effectiveness (PUE)

PUE is a representation of the ratio between the total power consumed by a data center *Pt* and the power used by the computer servers *Ps* i.e.,

$$PUE=Pt/Ps. \quad (1)$$

The non-computing—i.e. infrastructure—systems in a data center account for the difference between *Pt* and *Ps*. This includes the power consumed by HVAC systems and power losses due to the electrical resistance of existing carrying conductors, and the inefficiency of components in the electrical distribution system, which typically includes transformers, switchboards, and (PDU) Power Distribution Units adding up to ATS and UPS devices.

Figure 1 conveys an example power flow diagram for a data center to better convey power transformations and sources for losses and inefficiencies. A PUE of 1 means that the power utilized by the servers account for all of the power delivered to a data center. Since the purpose of a data center is to house these servers, a PUE of 1 is taken to represent a 100% competent data center from a power usage perspective. Also, it is impossible to have a PUE less than 1.

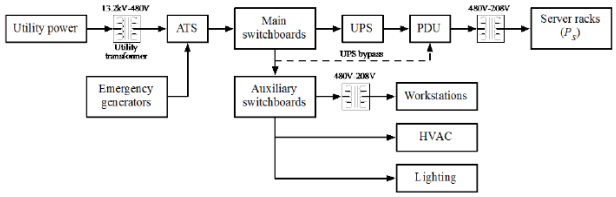

Figure 1 An example power flow diagram for a data center

D. Determinations Issues

There are several issues or challenges associated with PUE determinations and using the metric and a means of comparing data center facilities. One of the more elementary problems is inconsistency when determining what infrastructure components constitute *Pt*. Some consider *Pt* to represent the "useful" power entering a data center that specifically supports the continued operation of the computer servers. There are several other issues associated with PUE. For instance, IT specialists may choose to implement virtualization during data center operation as a means of reducing total power expenditure via decreased server utilization (Niles, 2008). Such activities, however, reduce Ps and therefore increase PUE, since the power used by the infrastructure components likely do not reduces proportionally with Ps [5].

E. Basic data centre metrics

These metrics have been used for the design, reflecting the capacity attributes for a data centre and don't express the overall availability and efficiency:

(a) Power/area (W/m2) = power density for the data centre but not reflect the sizing of specific cooling or power architecture. The layout of computer room – the density of IT racks – is a determinant factor. Usually an IT rack has 0.8 m2 and has allocated an average of 3.2 m2 space in the computer room. Under this value is considered high density design that leads to increase power.

(b) Power/rack (kW/rack) = power density metric reflects the distribution of equipment on surface. Racks with 2 kW or 20 kW can be in the same space and in close proximity to each other. Providing power and cooling to these racks may require different solutions. For standard design dimensioning, the load is considered maximum 7 kW/rack and an average of 5 kW/rack. Also the power consumed by the rack is according with IT hardware technology and software applications.

(c) Cost/area (€/m2) = the data centre power density can vary greatly and total square footage is independent of this value; as such will lead to poor value. The metric is more useful to estimate general office or industrial space cost and less for data centre. The cost/m2 of data centre is also according with quality of services, redundancy and availability of associated facilities (power, cooling, mechanics, security etc.). The estimated cost vary from 3800 €/m2 Tier I to +8000 €/m2 for Tier IV data centre .

(d) Cost/power (€/kW) = the main data centre cost is in the power and cooling infrastructure. Therefore the kW cost for connecting IT load is a parameter for interpreting the layout efficiency and rack room power density [7].

F. Advanced Configuration and Power Interface (ACPI)

An open industry standard, permits an operating system to directly control the power-saving aspects of its underlying hardware. Some programs allow users to adjust the voltages supplied to the CPU through a process called under volting , which reduces both the amount of electricity consumed and heat produced. Some CPUs can automatically under volt the processor, depending on the workload. Such a technology is called "SpeedStep" on Intel processors, "PowerNow!" and "Cool'n'Quiet" on AMD chips, LongHaul on VIA CPUs, and LongRun on Transmeta processors. The technique of workload dependent dynamic power management refers to dynamic core power and speed adjustment according to the current workload, i.e., the number of applications in a server system, the distribution of the applications among the cores, and the characteristics of the applications. For instance, the power supply and the core pace are increased when there are more tasks in a server, like that tasks can be processed faster and the average task response time is reduced. On the other hand, the power supply and the core speed are decreased when there are less tasks in a server, like that energy consumption can be reduced without significant performance degradation. Such runtime power and speed adjustment can be supported by a mechanism named dynamic voltage scaling, or equivalently, dynamic frequency scaling, dynamic speed scaling, or dynamic power scaling [6].

G. Data Center Architecture

The most widely used data center architecture is in the form of three-tier trees of host servers and switches. This architecture (Fig. 2) consists of the core tier at the root of the tree, the aggregation tier that is accountable for routing, and the access tier that holds the pool of computing servers (or hosts). Compared with the early two-tier data centers that supported no more than 5,000 hosts, the three-tier infrastructure can contain much more computing servers, e.g., a DC with the capacity of 100,000 hosts.

H. Data Center Network Traffic

Many distributed software systems, such as MapReduce and scientific applications running in DCs spawn a large number of tasks with complex communication. These tasks often process chunks of large dataset and communicate with each other. Thus, DCNs need to carry much more traffic than ever before [8].

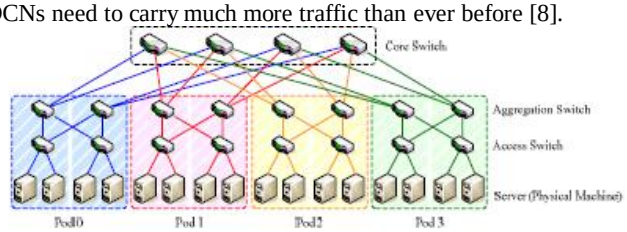

Figure 2 Fat-tree architecture with 4-port switches

II. LITERATURE SURVEY

Chao-Tung Yang et al. [2011] With the development of electronic of government and business, the implementation of these services are increasing the demand for servers, each year a considerable amount of the procurement server and out of the server are too old to provide better service. However, due to the speed of the server out of nowhere near the rate of increase, the continued development of the server, on behalf of our need to prepare more space, power, air conditioning, network, human and other infrastructure. Derived from these costs, long years, the often less than the purchase price of the server. And the provision of these services is essentially quite energy-intensive, especially when the server is running at low utilization, the making idle resources, waste, which is caused by the

power efficiency of data centers the main reason for the low. Even in a very low load, such as 10% CPU utilization, the total power consumption is more than 50% in the peak. Similarly, if the disk, network, or any that resource is the bottleneck, it will increase the waste of other resources. The "Green" became a hot key word recently. And intended the topic and proposed power management approach with virtualization technology.

Saurabh Kumar Garg et al. [2011] The energy efficiency of ICT has become a major issue with the increasing demand of Cloud Computing. More and more companies are investing in building large datacenters to host Cloud services. These datacenters not only consume huge amount of energy but are also very difficult in the infrastructure itself. Many studies have been proposed to make these datacenter energy efficient using technologies such as virtualization and consolidation. Still, these solutions are generally cost driven and thus, do not directly address the critical impact on the environmental sustainability in terms of CO2 emissions. Hence, in this work, propose a user-oriented Cloud architectural framework, i.e. Carbon Aware Green Cloud Architecture, which addresses this environmental problem from the overall usage of Cloud Computing resources. Also present a case study on IaaS providers. Finally, we present future research directions to enable the wholesome carbon efficiency of Cloud Computing.

Nosayba El-Sayed et al.[2012] The energy consumed by data centers is starting to make up a significant fraction of the world's energy consumption and carbon emissions. A large fraction of the consumed energy is used up on data center cooling, which has motivated a large body of work on temperature management in data centers. Interestingly, a key aspect of temperature management has not been well understood: calculating the setpoint temperature at which to run a data center's cooling scheme. Most data centers set their thermostat based on (conservative) suggestions by manufacturers, as there is restricted understanding of how higher temperatures will affect the system. At the same time, studies suggest that increasing the temperature setpoint by just one degree could save 2–5% of the energy consumption. This paper gives a multi-faceted study of temperature management in data centers. Author use a large collection of field data from different production environments to study the effect of temperature on hardware reliability, including the reliability of the storage subsystem, the memory subsystem and server reliability as a whole. Author also use an experimental testbed based on a thermal chamber and a great array of benchmarks to study two other potential issues with higher data center temperatures: the effect on server performance and power. Based on conclusion, author make recommendations for temperature management in data centers that create the potential for saving energy, while limiting negative effects on system reliability and performance.

Marina Zapater et al. [2013] The computational and cooling power demands of enterprise servers are increasing at an unsustainable rate. Understanding the relationship between computational power, temperature, leakage, and cooling power is critical to enable energy-efficient operation at the server and data center levels. This paper develops empirical models to estimate the contributions of static and dynamic power expenditure in enterprise servers for a wide range of workloads, and analyzes the interactions between temperature, leakage, and cooling power for various workload allocation policies. Author proposes a cooling management policy that minimizes the server energy consumption by setting the optimum fan speed during runtime. Experimental results on a currently shipping enterprise server demonstrate that including leakage awareness in workload and cooling management provides additional energy savings without any impact on performance.

Jumie Yuventi et al. [2013] Data centers represent an increasingly popular construction project type, supported by the continued growth in internet-based services. These facilities can, however, consume large amounts of electricity and—particularly if growth trends continue—put strain on utility grids and energy resources. Many metrics have been proposed to evaluate and correspond energy use in data centers. In many cases, the goal is that these metrics will be used to develop energy conscious behavior and perhaps data center sustainability ratings or building codes to reduce average energy use. In this paper, author examine one of the more popular metrics, Power Usage Effectiveness (PUE), and discuss its shortcomings toward effectively communicating energy sustainability. Author assumption is that PUE is an instantaneous representation of electrical energy consumption that encourages operators to report the minimum observed values of PUE. Hence, PUE only conveys an understanding of the minimum probable energy use. Instead author propose the use of energy-based metrics or average PUE over a significant time period—e.g., a year—to better comprehend the energy efficiency of a data center and to develop sustainability rating/ranking systems and energy codes.

Keqin Li et al. [2015] The technique of using workload dependent dynamic power management (i.e., changeable power and speed of processor cores according to the present workload) to improve system performance and to reduce energy consumption is investigated. Typically, the power supply and the core pace are increased when there are more tasks in a server, such that tasks can be processed faster and the average task response time is reduced. On the other hand, the power supply and the core speed are minimized when there are less tasks in a server, such that energy consumption can be reduced without significant performance degradation. A queueing model of multicore server processors with workstress dependent vibrant power management is established. Several speed schemes are proposed and it is demonstrated that for the same average power consumption, it is probable to design a multicore server processor with workload dependent dynamic power management, such that its average task response time is shorter than a multicore server processor of regular speed (i.e., without workload dependent dynamic power management). It is shown that given certain application environment and average power consumption, there is an optimal speed scheme that decreases the average task response time. For two-speed schemes, the problem of optimal design of a two-speed scheme for given power supply and power utilization model is formulated and solved. It is pointed out that power consumption reduction subject to performance restrictions can be studied in a similar way as performance improvement (i.e., average task response time reduction) subject to power expenditure constraints. To the best of our knowledge, this is the first work on analytical study of workload dependent dynamic power management.

Cătălin Dumitrescu et al. [2016] This paper proposes alternatives of energy efficiency assessment for data centres. Data centres are the core of present information and communication economy. Since the early days of telecom's facilities, the operators' main challenges were the equipment placement, capacity planning, expansion, redundancy and availability of services, all these factors leading to power consumption inefficiencies. The evolution of Information and Communication Technology (ICT or IT&C) and cloud computing, increase the storage and processing capacity conduct to an exponential growth of Information Technology (IT) resources that have to be hosted and provide continuing operation. Beside of availability factor, the optimal utilization of the capacity and energy efficiency are the main characteristics of Data Centre facility. The overall efficiency of a data centre has to be precisely calculated to establish parameters of comparison, estimate environmental impact and to decide the specific measures for reducing energy consumption. The practical design, simulation and measurement

have made it possible a comparison of efficiency parameters regarding energy consumption and IT utilization density.

Ting Yang et al. [2014] Data Center (DC), the underlying infrastructure of cloud computing, becomes startling large with more powerful computing and communication capability to assure the wide spectrum of composite applications. In a large scale DC, a great number of switches connect servers into one complex network. The power consumption of this communication network has skyrocketed and become the same league as the computing servers' costs. More than one-third of the total energy in DCs is frenzied by communication links, switching and aggregation elements. Saving Data Center Network (DCN) energy to improve data center efficiency (power usage effectiveness or PUE) become the key technique in green computing. In this paper, author present VPTCA as an energy-efficient data center network planning solution that collectively deals with virtual machine placement and communication traffic configuration. VPTCA aims to lessen the DCN's energy consumption. In particular, interrelated VMs are assigned into the same server or pod, which effectively helps to reduce the amount of transmission stack. In the layer of traffic message, VPTCA optimally uses switch ports and link bandwidth to balance the load and avoid congestions, enabling DCN to increase its transmission capacity, and saving a considerable amount of network energy. In our evaluation via NS-2 simulations, the performance of VPTCA is measured and compared with two well-known DCN management algorithms, Global First Fit and ElasticTree. Based on experimental results, VPTCA outperforms existing algorithms in providing DCN more transmission capacity with less energy consumption.

Table 1 Summary of various techniques

Sr. No	Year	Tech. used	Outcome
1	2011	Green Power Manager Algorithm	GPM approach got a signification energy saving than traditional approach.
2	2011	Carbon Aware Green Cloud Architecture	Green Policy CEGP can save up to 23% energy while improving the carbon footprint by about 25%.
3	2012	Temperature Management in Data Centers	Most organizations could run their data centers hotter than they currently are without making significant sacrifices in system reliability.
4	2013	Using power model leakage-aware cooling control policy applies	optimizing CPU power consumption by 2.5% for the whole cluster, and showing how the impact of policy raises as data room temperature increases.
5	2013	Adjustment of PUE that is based on energy readings or estimates. This adjusted metric is "Energy Usage Effectiveness" (EUE)	EUE is directly communicating energy efficiencies. Also, it may be possible to determine EUE directly for electrical bills
6	2015	several speed schemes	given certain application environment and average power consumption, there is an optimal speed scheme that minimizes the average task response time.
7	2016	Energy Efficiency Parameters in Data Centres	The external factors that influence the overall performance and efficiency for a facility are: Climate and Temperature which have big impact on cooling efficiency, and the internal ones are the Technology and Infrastructure utilization.
8	2014	VPTCA as a novel DCN planning solution, incorporating a GA-based VM placement algorithm and a traffic configuration algorithm.	VPTCA outperforms other two well known traffic assignment algorithms in in providing more transmission capability with less energy consumption.

III. Conclusion

Even though Cloud Computing is a great innovation in the world of computing, there also exist downsides of cloud computing. Due to scale and complexity of data center equipment it is extremely difficult to define unique service or activity that could be examined for its energy efficiency. Except using Cloud Computing concept for its main purpose, the Cloud Computing infrastructure and its flexible nature can also be utilized indirectly for energy

optimization. Techniques such as DVFS and powering on/off machines can be used for frequency regulation of a power network. Using this technique, data center's dynamic load can be used for regulating electricity demand and thus production, which is important for keeping optimal frequency of the power grid. In order to achieve significant improvements, energy efficient solutions have to cover individual components, as well as their integration with the rest of the system. Great energy savings can be achieved by turning more servers into lesser power states and by increasing the utilization of the already active ones. Three approaches to achieve savings are: workload prediction, VM placement and workload consolidation, and resource over commitment.

References

1.Chao-Tung Yang, Kuan-Chieh Wang, Hsiang-Yao Cheng, Cheng-Ta Kuo1, and Ching-Hsien Hsu, "Implementation of a Green Power Management Algorithm for Virtual Machines on Cloud Computing", Springer-Verlag Berlin Heidelberg , pp. 280–294, 2011.

2. Saurabh Kumar Garg, Chee Shin Yeo and Rajkumar Buyya, "Green Cloud Framework For Improving Carbon Efficiency of Clouds", Springer-Verlag Berlin Heidelberg , pp. 1-13, 2011

3. Nosayba El-Sayed , Ioan Stefanovici, George Amvrosiadis ,Andy A. Hwang and Bianca Schroeder, "Temperature Management in Data Centers: Why Some (Might) Like It Hot**",** *SIGMETRICS'12,* pp.1-12, 2012

4.Marina Zapater, Ozan Tuncer, José L. Ayala, José M. Moya, Kalyan Vaidyanathan, Kenny Gross and Ayse K. Coskun, " Leakage-Aware Cooling Management for Improving Server Energy Efficiency", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, pp.1-14, 2013

5. Jumie Yuventi, P.E. and Roshan Mehdizadeh, "A critical analysis of Power Usage Effectiveness and its use as data center energy sustainability metric", Civil & Environmental Engineering, Stanford University, Stanford, CA 94305, pp.1-11, 2013.

6. Keqin Li,, "Improving Multicore Server Performance and Reducing Energy Consumption by Workload Dependent Dynamic Power Management", IEEE TRANSACTIONS ON CLOUD COMPUTING, VOL. XX, NO. YY, MONTH, pp.1-14, 2015.

7. Cătălin Dumitrescu, Adrian Pleșca, "Overview on Energy Efficiency Parameters in Data Centres", International Conference and Exposition on Electrical and Power Engineering (EPE 2016), pp.153-156.

8.Ting Yang, Young Choon Lee ,Albert Y. Zomaya, " Energy-Efficient Data Center Networks Planning with Virtual Machine Placement and Traffic Configuration", IEEE 6th International Conference on Cloud Computing Technology and Science, pp. 284-291, 2014

First Author Jagjeet, M.Tech. Scholar.
Second Author Manju Ahuja, Assistant Professor.