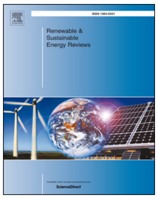
Renewable and Sustainable Energy Reviews 94 (2018) 363–385

Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

journal homepage: www.elsevier.com/locate/rser



Energy, performance and cost efficient datacenters: A survey[☆]

Muhammad Zakarya[a,b]

[a] *Department of Computer Science, University of Surrey, UK*
[b] *Abdul Wali Khan University, Mardan, Pakistan*

ARTICLE INFO

Keywords:
Datacenters
Resource management
Server consolidation
Dynamic capacity planing
VM migration
Energy efficiency

ABSTRACT

Computing systems have been focused on performance improvements, driven by the demand of user applications in past few decades, particularly from 1990 to 2010. However, due to their ever-increasing energy demand which causes large energy bills and CO_2 emissions, over the past six years the focus has shifted towards energy-performance aware. The average energy consumption of servers is increasing continuously; and several researchers suggest, if this trend continues further, the cost of energy consumed by a server during its lifetime will exceed the hardware costs. The energy consumption problem is even greater for large-scale infrastructures, such as clusters, grids and clouds, which consist of several thousand heterogeneous servers. Efforts are continuously made to minimize the energy consumption of these systems, but the interest of people in computational services and popularity of smart devices make it a difficult task. In this paper, we discuss the energy consumption of ICT equipment, and present a taxonomy of energy and performance efficient techniques for large computing systems covering clusters, grids and clouds (datacenters). We discuss both energy and performance efficiency, which makes this survey different from those already published in the literature. Key research papers are surveyed and mapped onto taxonomies to characterise and identify outstanding and key issues for further research. We discuss several state-of-the-art resource management techniques, reported in the literature, that claim significant improvement in the energy efficiency and performance of ICT equipment and large-scale computing systems such as datacenters, and identify a few open challenges.

1. Introduction

Large-scale computing systems as observed in the top500 [1] supercomputers, clusters, grids and clouds [2], consist of many Information and Communication Technology (ICT) devices that are inter-connected through computer networks. Supercomputers and clusters are examples of non-distributed systems that are largely used to run low latency services. Furthermore, these systems are also used to solve big problems quickly where bulky mathematical operations and computations are involved, such as defense and control systems, weather forecasting etc. Distributed systems, such as grids and clouds, might be preferred over non-distributed systems for several reasons including concurrent execution, reliability and distributed nature of applications [2]. These systems offer their services to customers based on either commercially reasonable or best effort policies. Furthermore, grids and clouds may be distributed over different locations, connected over networks. Therefore, these systems are not preferred for low latency services. With the development of High Performance Computing (HPC) clouds and Graphics Processing Unit (GPU) servers inside datacenters, processing can be increased. However, performance of low latency services will remain an issue for underlying networks. There are certain distributed systems such as HPC, that provide low latency services with high compute performance.

Cluster, grid and cloud service providers maintain shared pools (of different capacities) of computational resources (e.g. servers, storage) known as datacenters. Each datacenter needs energy to: (i) operate properly, and (ii) cool-down the heat produced by servers. In 2010, the energy consumption by datacenters was projected to be in the range of 1.1–1.5% of the worldwide energy use and is expected to increase further in the near future. Datacenters are the principal electricity consumers in cloud computing, reportedly consuming approximately 70 billion kWh in 2014, equivalent to 1.8% of the US total energy consumption, and are projected to account for approximately 73 billion kWh by 2020 [3]. It has been suggested that due to resource management techniques like virtualisation and consolidation of Virtual Machines (VMs) [4] this figure (~ 70 billion kWh) increased by only 4% from 2010 to 2014, which is a significant improvement over the 24% increase from 2005 to 2010 [5]. Industry reviews such as [3,6] have

[☆] This paper is an extension of our work, published in Sustainable Computing: Informatics and Systems, Elsevier (2017). A very small part is also taken from our work which was presented in the 13th International Conference on Economics of Grids, Cloud, Systems and Services and appeared in Lecture Notes in Computer Science, Springer (2016).

E-mail address: mohd.zakarya@surrey.ac.uk.

https://doi.org/10.1016/j.rser.2018.06.005
Received 17 September 2017; Received in revised form 13 May 2018; Accepted 1 June 2018

illustrated that US datacenters (including small and large clusters) energy cost had increased by 15% per year since 2011.

A report [7] describes that in 2016, all over the world, nearly ~ 416.2 billion kWh of energy was consumed by compute servers, which is even more than the total energy consumption of the UK. Energy consumption will continue to rise further with growing demand and capacity of datacenters' resources unless energy efficient management techniques such as scheduling, resource allocation and management algorithms are established, designed and applied [8]. The resource scheduling and management algorithms, together with the physical infrastructure (resources, set-up, location etc.) of the clusters, grids and clouds are supposed to decrease the ecological (environmental) impact (CO_2 emissions) and make these systems highly energy, performance and cost efficient [9].

Both environmental and economical issues related to large-scale datacenters encourage us to complete this survey. Recently, due to the quick uptake of cloud providers to run industrial applications, minimizing the operational costs of large-scale datacenters (powering and cooling), such that the application performance is adversely not affected, is a key economical concern. For every 10 °C increase in server's temperature, its failure rate almost doubles, therefore, least temperature may mean improvements in datacenter reliability as well [10]. Many studies are being completed that investigate and elaborate on energy-aware computing and datacenters. The goal of this survey is to analyse the energy consumption of ICT equipment, in general, and focus on large-scale systems such as clusters, grids, clouds and, particularly, datacenters. Even though the scope of this survey is limited to those techniques that claim energy efficiency at scheduling and resource management levels. Nevertheless, where essential, other methods are also briefly described. We describe a taxonomy of methods that are suggested to improve the energy and performance efficiency of large-scale systems. Theoretically, clusters, grids and clouds are treated as identical [2], therefore, significant efforts have been made to analyse and differentiate the energy and performance aware resource management methods suggested for these systems. The major contributions of the survey are as follows:

1. a taxonomy of energy and performance efficient resource management techniques in cloud datacenters;
2. system-level energy efficient scheduling (CPU) for individual systems;
3. cluster-level scheduling techniques to reduce the energy consumption of computational clusters;
4. datacenter-level resource management for energy and performance efficiency in computational grids and virtualised clouds; and
5. energy and performance efficiency is described for different kinds of workloads.

In this survey, we describe the energy consumption of computational devices (individual servers), clusters, grids and clouds, and look for possibilities to reduce their energy consumption without adverse impact on service performance and quality. Our survey is different from those conducted in [5,11–14]. In this survey, we extend our own survey conducted in [5] in order to account for: (i) performance aware computing; (ii) energy-performance efficient datacenters; and (iii) new findings and directions for future research. The techniques demonstrated in [11], offer a taxonomy of optimizations, but the energy efficiency methods in virtualised cloud environments have not been investigated. Similarly, the surveys completed in [12,13], focus on energy efficient datacenters, but energy efficiency of system-level (CPU) scheduling techniques is not discussed. Furthermore, performance of cloud's workload is not explored. We start from describing the efficiency of a single system, its components and explore efficiency of large-scale systems, particularly datacenters, in terms of energy consumption and performance. We believe that this survey will offer readers an understanding of the most important ideas in the field of energy, performance and cost efficient resource management techniques in large-scale systems, particularly, datacenters. Moreover, this survey will assist scholars to identify the outstanding and key issues for further research.

Table 1
Worldwide datacenters energy consumption in billion kWh (2000–2016) [15],– for 2010, the energy usage has also been reported in the range of 271.8–301.1b kWh.

Year	2000	2005	2010	2016
Energy consumption	70.8	152.5	397.6 271.8–301.1	416.2

The rest of the paper is organized as follows. In Section 2, we describe the energy consumption problem. An overview of the background study is presented in Section 3. To decrease servers' power consumption, two kinds of power management techniques are illustrated in Section 4. Section 5 presents a taxonomy of energy and performance efficient resource management techniques at datacenter level. It is also essential to measure the energy consumption of ICT equipment such as server, when dealing with energy related issues. Therefore, this section also presents mathematical models to estimate the energy consumption of a virtualised host and VM migration. Key findings of this survey are explained in Section 6. In Section 7, a gap has been identified for further research in the domain of energy and performance efficient datacenters. Section 8 describes several metrics that service providers use to measure energy efficiency of their infrastructures (datacenters). Finally, we summarise the findings of this survey in Section 9.

2. The energy consumption problem

The large number of hosts inside datacenters consume energy and produce Greenhouse Gas (GHG) emissions. Energy and fuel cost are rising that could also affect the economy of service providers. If we assume a typical compute server, which is consuming ~ 450 Wh, and the US commercial electricity price of 0.08$ per kWh, it will cost ~ 315$ per year if it is fully utilized over the year. This figure would translate into a million dollar energy bills for clusters consisting of more than three thousand servers. Therefore, to minimize the energy bills that will also maximize the provider's profit and will minimize GHGs emissions, it is important to consider the energy efficiency of these large-scale systems. Table 1 shows worldwide datacenters energy consumption from 2000 to 2016 [15][1], which is continuously increasing.

Kaur et al. [13] projected that the global energy sector is responsible to produce approximately 43% of the total GHGs. Following the principles of energy-aware computing, it is necessary to decrease datacenters energy consumption that should subsequently reduce GHG emissions. In 2007, the IT sector (including servers, cooling equipments, PCs, networks, telephony, printers, mobiles and office telecommunication) energy requirements and GHG emissions were considered almost equal to those of airline industry, which is 2% of the worldwide GHG emissions [16,17]. In 2013, the US datacenters and servers consumed approximately 91 billion kWh of energy and were estimated to be consuming nearly 140b kWh annually by 2020 [18]. Nevertheless, due to state-of-the-art energy efficiency methods, a most recent study [3] suggests that these systems currently account for approximately 70b kWh of energy, which is 1.8% of the US' total energy consumption and are projected to consume roughly 73b kWh by 2020. Moreover, the same study suggests the current share of ICT equipment approximately 1.6% to worldwide GHG emissions and predicts this figure to be almost 2% by 2020 [19]. Table 2 shows the percentage of

[1] http://www.independent.co.uk/environment/global-warming-data-centres-to-consume-three-times-as-much-energy-in-next-decade-experts-warn-a6830086.html.