Computers in Industry 65 (2014) 401–407

Contents lists available at ScienceDirect

Computers in Industry

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

Sustainability in Web server systems

Jose M. Sola-Morena [a,*], Katja Gilly [a], Carlos Juiz [b]

[a] Física y Arquitectura de Computadores, Miguel Hernández University, Elche, Spain
[b] Departament de Ciències Matemàtiques i Informàtica, University of Balearic Islands, Palma de Mallorca, Spain

ARTICLE INFO

Article history:
Received 28 February 2013
Received in revised form 29 July 2013
Accepted 8 November 2013
Available online 4 December 2013

Keywords:
Energy efficiency
Green networking
Web servers
Load balancing
Power consumption

ABSTRACT

The exponential growth of Internet during the last decade leads us to make more efforts in the researching and developing of sustainable Web servers in order to decrease the global energy demand. In this paper, we cover, as a first step, a proper review of the literature related to the energy efficiency research in Web server systems to depict the state of the art to plan further contributions as more research in sustainable Web systems. We also propose and implement an energy metric that permits to establish a relation between the Quality of Service (QoS) obtained by the system and the power it consumes.

1. Introduction

The global increase of Internet in the last ten years has provided a large new set of user-oriented services. However, these new high quality services imply a higher level of power consumption. This is becoming an important concern from the environmental point of view but also as a critical issue in terms of economical investment for the Internet service providers. Starting from this need, and taking into consideration that these new high quality Internet services coming soon in future require not only a higher consumption, but a larger number of devices. Hence, new methods to reduce energy consumption are needed. Some of these methods and approaches are called "green networking" and will be described later.

It is shown in [1] how silicon technologies (e.g. Complementary Metal Oxide Semiconductor (CMOS)) improve their energy efficiency with a lower rate with respect to router's capacities and traffic volumes, by increasing a factor of 1.65 every 18 months, while the Internet data traffic volume increase follows Moore's law, by doubling every 18 months. These facts together represent a good example of the power deficit in energy consumption on Internet services.

The European Commission Directorate-General for Information Society and Media (DG INFSO) report estimated that the total Internet energy consumption will double in the next ten years [1,2]. This energy consumption estimation shows the energy consumption (in TWh) for the European Telecommunication's network infrastructures in Business as Usual (BAU) and in Eco Sustainable (ECO) scenarios, as well as the cumulative energy savings between the two scenarios. BAU is based on the assumption that, for the next ten years, no new energy saving method will be implemented, while for the ECO scenario, the estimation for the next years takes into consideration that network infrastructures will envisage the energetic problem by implementing energy saving methods. In that sense, the predicted cumulative energy savings show an important difference in terms of the total energy saved (in TWh) between using ECO and BAU networks.

Energy efficiency improvement is then becoming a critical research item since environmental and economic constraints surround the exponential growth that Internet has been experiencing during last years. Hence, the energy consumption increase brings two problems, such as:

1. The environmental concern, measured by the amount of carbon dioxide (CO_2) emissions. This is basically the reason why the energy concern in network systems is usually called the "green networking".
2. The economic problem, mainly affecting Internet service companies, who may keep a good level of QoS, even taking into consideration the increase of high quality services demand (such a multimedia/broadcasting high quality services), and try to keep costs as low as possible.

* Corresponding author.
E-mail address: jose-m.sola@graduado.umh.es (J.M. Sola-Morena).

0166-3615/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.compind.2013.11.009

Concerning the environmental problem, the growth of ICT industry implies a necessary increase of CO_2 emission due to the need of more hardware infrastructure, and hence, more cooling equipment, that is constantly consuming energy (approximately the 2% of global CO_2 emissions, overcoming even the carbon footprint of aviation [1]). In detail, focusing on telecommunication networks, it is estimated to produce about 1.4% of the global CO_2 emissions in 2020 (being now less than 1%).

Therefore, fixed and mobile network infrastructures have enormous and heavily increasing requirements in terms of electrical energy. Especially in Internet, where new high quality services need more components in order to provide a good QoS for users and hence, more power. This increase in the demand of energy implies both an environmental and economic impact.

The purpose of this paper is to give a proper overview and state of the art of the subject, as first step for future contributions in next-generation methods and scenarios, based on this preliminary research of existing methods. In a second part, we perform and simulate different scenarios as well as test several parameters influencing Web energy consumption. A preliminary version of this paper was already presented in [3].

The rest of this paper is organised as follows: in Section 2, current energy aware Web proposals are mentioned, reviewed and compared. Based on some of the works considered, we have designed some simulation scenarios that are described in Section 3, explaining which assumptions have been made for their definition. Finally, in Section 4, we present the results obtained to conclude with some discussion, concluding points and open problems.

2. State of the art

The first part of this paper intends to be a survey of existing proposals and trends related to energy efficiency in Web server systems in order to make them more sustainable.

The first relevant approach in energy savings was introduced by Pinheiro et al. [4]. This paper states that most of the time servers located in a server cluster are consuming energy while they are not even being used. Hence, the technique of turning off and on servers, depending on the instantaneous needs of the system, is introduced making the cluster dynamic (since turning off a server means to remove a node from the system, in the same way that turning a server on means to add a node). Pinheiro et al. do not completely disregard prediction, but use a basic approach: prediction based on "tracking" the past and current traffic on the system. The results obtained when testing the method with an eight nodes server cluster are optimistic: for high loads, the energy saving is relatively small, while for low loads, the energy saving percentage (compared to the same system with no optimisation) is up to almost 80% of power saving. In this work, a comparison in terms of power consumption between static and dynamic Web cluster configuration is shown, proving that the performance obtained is much better on the dynamic one in terms of energy saving.

The method presented in [4] has two drawbacks: the prediction method is very basic and the amount of time that needs the servers to turn off and on is long (around 45 s and 100 s, respectively).

Another relevant energy saving method is the Dynamic Voltage Scaling (DVS) where the voltage of the server's processor is increased or decreased depending on a different factor, i.e. on the load sent to this server. This method was presented in [5] by Horvarth et al. Applying two different types of load on the tests the results show a reduction of around 30% of power consumption.

The higher the load is, the lower the energy saving performance obtained. Hence, again the maximum savings (in terms of %) are obtained for low workloads. It is important to mention that in [5], the frequency for the CPU speed adjustments change depending on the load, which means that for higher loads, the CPU adjustment takes place more often.

Three different DVS methods are compared in [5]: the independent DVS (normal one), feedback DVS (with adaptive behaviour) and weighted DVS. In this paper, the authors provide data about the power savings (in percentage) using those three methods, comparing them with a scenario where no power saving method at all was performed. As it will be described later in this paper, several authors consider using DVS combined with other power saving methods.

In [6], Bertini et al. present a model that consists of a front-end server that redirects the incoming load from the users (in this case, Web users) to the corresponding server. This means that they implement the so-called "load balancing", i.e. redirect the load to the best server for each client request. These servers are connected to databases.

The power optimisation policies that the authors use in this paper are based on the ON/OFF switching and the DVS method. The result presented is a power performance able to maintain a good QoS with up to four times less power consumption than an equivalent scenario with no power saving methods implemented. No load prediction method is considered.

Like in the previous model, a load balancer is introduced in [7]. However, in this case, a Peer to Peer (P2P) network is considered, which means that each peer computer in the system can play both roles (server and client). Enokido et al. presented three different types of load balancing methods in [7]: Round Robin, Computation Laxity Based (CLB), which basically works calculating how long it will take for a server to perform an action and chooses the best one, and Power Consumption Laxity Based (PCLB), which assigns the different incoming tasks to the best server in terms of energy saving criteria. The results showed that PCLB obtains the lowest power consumption.

Santana et al. [8] also implemented load balancing on a similar scenario to the one used in [6] by Bertini et al., but presenting some results comparing three types of prediction methods: Performance Governor (PG), which is a linux system approach that works with all the servers at the maximum frequency, i.e. there is no power management, On-demand Governor, which implements DVS depending on the load, and finally, Optimised ON/OFF (OOO), which is a forecasting method implementing ON/OFF servers switching and DVS. In [8], the authors compare the three mentioned approaches with respect to Holt Linear Method (HLM) load prediction method (the one that is finally considered as the best one).

Bolla et al. introduced load prediction methods in power management policies [2], implementing and analysing a performance test on software routers, and bringing the different components of the software router to different states (i.e. working frequencies). These frequencies are a discrete set of values that have to be common to all the components of the software router. Taking into account that the consumption of an idle core is around half of the consumption of an active one, he uses as the prediction method a "time" division consisting in taking different time slices on the same day based on statistics from the past. In the test performed and explained in this paper, four different frequencies were used, and the results showed that with this approach, an average of 30% of the power could be saved (as usual, for high incoming loads, the power consumption management becomes more difficult). They performed an analysis for Commercial Off-The-Shelf (COTS) software routers in terms of power management (that are performed automatically by the routers) taking several CPUs and comparing various parameters (throughput, power consumption, etc.). The authors provide a representation of the power consumption as a function of the throughput and the frequency of the CPU. A relevant information that can be extracted