Future Generation Computer Systems 57 (2016) 1–12

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Evaluating the cooling and computing energy demand of a datacentre with optimal server provisioning

Ricardo Lent*

University of Houston, TX, United States

HIGHLIGHTS

- Assessment of the temperature setpoint in datacentres' energy consumption.
- Evaluation of cooling and computing factors with optimal server provisioning.
- Temperature-dependent redundancy with specific job service level objectives.
- Energy, temperature, and computing performance measurements from real servers.

ARTICLE INFO

Article history:
Received 18 May 2015
Received in revised form
17 September 2015
Accepted 14 October 2015
Available online 3 November 2015

Keywords:
Green computing
Datacentre
Servers
Performance and reliability
Energy measurements

ABSTRACT

As cloud computing continues to gain significance across fields, the energy consumption of datacentres creates new challenges in the design and operation of computer systems, with cooling remaining a key part of the total energy expenditure. We investigate the implications of increasing the room temperature setpoint in datacentres to save energy. For this, we develop a holistic model for the energy consumption of the server room that depends on user workload and service level agreement constraints, and that considers both cooling and computing energy dissipation. The model is applicable to a steady-state analysis of the system and brings insight into the impact of the most relevant parameters that affect the net energy consumption, such as the outside temperature, room temperature setpoint, and user demand. We analyse both static and dynamic server provisioning cases. In the latter case, a global power management scheme determines the optimal number of servers required to handle the incoming user demand to fulfil a target service level objective. Finally, we consider the extra energy needed to maintain service continuity under the expected higher server mortality rate due to warmer operational temperatures. Energy and temperature measurements acquired from a server machine running scientific benchmark programs allow to realistically fix model parameters for the study and to obtain pragmatic conclusions.

1. Introduction

Because of the high demand, cost and environmental reasons, the growing energy required to operate datacentres continues to be a serious problem. Industry studies have shown that datacentres are responsible for dissipating 1%–3% of the total global electrical energy. Moreover, a growing user demand for networked services, estimated at 30% per year [1], is driving a steady increase in datacentre energy consumption as larger workloads increase server utilisation, which in turn makes servers consume more energy. Furthermore, as user demand increases, the additional hardware needed to continue fulfilling service level agreements (SLA) with users pushes the energy consumption up. In addition to these factors, the rising electricity rates and the extra cooling costs due to global warming, create a clear need for higher energy efficiency in datacentres, and as a result, "green" management has become a topic of active research [2–4].

While some server manufacturers specify that inlet air temperatures may to go up to 35 °C without impacting the equipment, most operators rarely use high temperature setpoints because of the common belief that these temperatures may reduce the long-term reliability of hardware elements as the material of electronic components may be subject to irreversible changes under such conditions, likely affecting the system operation at some later time. Cooling is therefore used to maintain the operational temperature of servers within a recommended range that ensures the reliability

* Tel.: +1 7137434239.
E-mail address: rlent@uh.edu.

http://dx.doi.org/10.1016/j.future.2015.10.008
0167-739X/

of electronic components and therefore, of the services. The energy expenses due to the cooling infrastructure constitutes a significant fraction of a datacentre operating costs. Examining a particular case as an example [5], assuming a 460 m^2 datacentre, 52% of the total energy is consumed by the computing equipment, while 38% is used for cooling. That is, more than one third of the electricity bill can be due to cooling.

Despite recent studies have suggested that the impact of high datacentre temperatures on system reliability can be smaller than often presumed [6], the standard temperature setpoint across datacentres worldwide continues to be around 20 °C. In general, increasing temperature setpoints can help to reduce energy consumption. However, two factors may prevent energy reductions. The first factor is in the (grounded or not) higher risk of server failure. Redundant equipment can be added to reduce the perceived risks, but certainly, the extra equipment needed and the new operations involved (e.g., data copying) increases energy consumption. The second factor is that airflow inlet temperatures that are above 28 °C approximately make servers' fans spin faster, increasing the power use with the resulting increase in per-server energy consumption [7]. The extra power consumption at higher temperatures is also the result of the increase in leakage current of CMOS circuits. The threshold of 28 °C is not standard across server equipment and the extra power demand caused by servers' cooling fans may start as early as 25 °C (77 °F) in some other cases [7,6].

The energy consumption of a datacentre results from a complex interplay of different factors. This work attempts to fill a gap in the literature of a holistic analysis of energy consumption that considers both the energy dissipated due to the workload dynamics of computing equipment and cooling, as well as considering the possible side effects that could occur due to a change in the operating temperature. The contribution of this research is therefore, a new model of energy consumption for datacentres, that provides better insight into the tradeoffs of the temperature setpoint, the impact of external temperatures to the net energy consumption, and workload SLOs. The model is applicable to energy-performance studies of datacentre applications, such as cloud computing. An evaluation scenario using parameters obtained from measurements acquired on representative computing equipment demonstrates some of the tradeoffs and potential energy savings that could be achieved in a datacentre through temperature setpoint tuning.

2. Related works

With dynamic voltage and frequency scaling (DVFS), processors can modify their voltage and frequency parameters at runtime to optimise energy consumption [8,9] with the end objective of achieving energy proportionality [10]. As thermal issues become more significant to the management of dense computing deployments, a body of work has been directed towards achieving thermal efficiency. One approach has been the design of job scheduling algorithms for individual machines that can reduce average on-die temperatures [11]. The idea has been extended to achieve thermal distribution in datacentres [12,13]. Some of the techniques applied to achieve proper heat distribution include air flow control and thermal-aware scheduling [14], chaotic attractor predictors [15], and back-filling jobs to machines already allocated to run other jobs [16]. While these techniques might not lead to large reductions in energy consumption, the lower operating temperatures of servers allow increased reliability.

An estimation of the server temperatures can also aid in the optimisation of job distribution. The Mercury software suite provides single server temperature emulation from other system metrics, such as, utilisation, air flow, and heat flow [17]. ThermoCast provides a similar functionality but unlike Mercury, it includes the thermal effect of adjacent servers [18]. Other works have provided related energy and/or thermal models for simulation [19].

The heat produced by servers in the datacentre is transferred to the outside by a HVAC system, which normally uses chilled water cooled computer room air conditioning (CRAC) units and chillers that supply cold water to the CRAC units. Air blowers push cold air under the raised floor, which enter the room through vent tiles located in front of the racks. The hot air from the back of the racks returns to the CRAC units to be chilled [20]. The heat generated can then be ventilated or reused [21].

In this paper, we focus on evaluating energy efficiency and consider the macroscopic impact of heat transfers to electrical energy demand. We assume a simplified thermodynamic model that is adequate for a steady-state evaluation of the system. Detailed modelling of the heat transfer dynamics in the datacentre and temperature variations within the room can be found in the literature [22–26].

From a more practical perspective, a significant problem is how to fix the temperature setpoint of the room cooling system to achieve energy efficiency. As noted by Patterson [27], changing the temperature setpoint, which lies on the path between the ultimate heat source (the CPU) and the ultimate heat sink (the outside) may or may not improve energy efficiency given that the endpoints remain the same and higher operating temperatures can increase the power demand of some components in the datacentre. Despite, the analysis did not include the effect of the outside temperature, it is clear that increasing the temperature setpoint may not lead to significant energy savings given the complexities of the system. In fact, other studies suggested exploiting the diurnal patterns that provide different external temperatures to achieve energy savings by management cooling to include outside air and conditioned air via Direct Expansion units [28]. Similarly, temporal and geographic variations of the energy price have been used to minimise the total energy expenditure for distributed datacentres [29].

A crucial performance metric in datacentres is the response time, which is often linked with a Service Level Objective (SLO) giving a performance bound. The issue has received recent attention [30]. However, a holistic approach that examines the impact of the energy features of computing resources, a varying external temperature, the choice of the temperature setpoint, strategies for server provisioning with power management, and dependency on user workload with SLA constraints is still missing from the literature. In previous works, we studied the relationship between energy consumption and quality of service in terms of the jobs response time [31,32]. The following two works are perhaps the closest to ours. Li et al. [33] recently investigated the optimal load distribution in a datacentre considering an integration of both computing and cooling costs, but focusing on optimising air conditioning temperature unlike our work. Varsamopoulos et al. analysed the impact of energy proportionality on server provisioning in datacentres [34]. However, these works excluded the impact of the outside temperature. To our knowledge, the present work is the first to consider the outside temperature in addition to the temperature setpoint, user demand, dynamic server provisioning, and redundant hardware to maintain availability in an energy consumption model for datacentres.

3. Energy consumption model of a datacentre

We can classify the typical datacentre equipment into two main types: computing and support. The former type includes servers, networking equipment, and networked storage. The latter consists of equipment for power distribution and uninterruptible supply, cooling, and lighting. The power consumed by the majority of datacentre equipment depends on user workload and remains approximately constant for the rest. As a result, we can express the