Applied Soft Computing 49 (2016) 94–107

Contents lists available at ScienceDirect

Applied Soft Computing

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

Runtime data center temperature prediction using Grammatical Evolution techniques

Marina Zapater[a,c,b,*], José L. Risco-Martín[a], Patricia Arroba[c,b], José L. Ayala[a], José M. Moya[c,b], Román Hermida[a]

[a] DACYA, Universidad Complutense de Madrid, Madrid 28040, Spain
[b] CCS – Center for Computational Simulation, Campus de Montegancedo UPM, 28660, Spain
[c] LSI – Integrated Systems Lab., Universidad Politécnica de Madrid, Madrid 28040, Spain

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form 26 June 2016
Accepted 25 July 2016
Available online 16 August 2016

Keywords:
Temperature prediction
Data centers
Energy efficiency

ABSTRACT

Data Centers are huge power consumers, both because of the energy required for computation and the cooling needed to keep servers below thermal redlining. The most common technique to minimize cooling costs is increasing data room temperature. However, to avoid reliability issues, and to enhance energy efficiency, there is a need to predict the temperature attained by servers under variable cooling setups. Due to the complex thermal dynamics of data rooms, accurate runtime data center temperature prediction has remained as an important challenge. By using Grammatical Evolution techniques, this paper presents a methodology for the generation of temperature models for data centers and the runtime prediction of CPU and inlet temperature under variable cooling setups. As opposed to time costly Computational Fluid Dynamics techniques, our models do not need specific knowledge about the problem, can be used in arbitrary data centers, re-trained if conditions change and have negligible overhead during runtime prediction. Our models have been trained and tested by using traces from real Data Center scenarios. Our results show how we can fully predict the temperature of the servers in a data rooms, with prediction errors below 2 °C and 0.5 °C in CPU and server inlet temperature respectively.

1. Introduction

Data Centers are found in every sector of the economy and provide the computational infrastructure to support a wide range of applications, from traditional applications to High-Performance Computing or Cloud services. Over the past decade, both the computational capacity of data centers and the number of these facilities have increased tremendously without relative and proportional energy efficiency, leading to unsustainable costs [1]. In 2010, data center electricity represented 1.3% of all the electricity use in the world, and 2% of all electricity use in the US [2]. In year 2012, global data center power consumption increased to 38GW, and in year 2013 there was a further rise of 17% to 43 GW [3].

The cooling needed to keep the servers within reliable thermal operating conditions is one of the major contributors to data center power consumption, and accounts for over 30% of the electricity bill [4] in traditional air-cooled infrastructures. In the last years, both industry and academia have devoted significant effort to decrease the cooling power, increasing data center Power Usage Effectiveness (PUE), defined as the ratio between total facility power and IT power. According to a report by the Uptime Institute, average PUE improved from 2.5 in 2007 to 1.65 in 2013 [5], mainly due to more efficient cooling systems and higher data room ambient temperatures.

However, increased room temperatures reduce the safety margins to CPU thermal redlining and may cause potential reliability problems. To avoid server shutdown, the maximum CPU temperature limits the minimum cooling. The key question of how to set the supply temperature of the cooling system to ensure the worst-case scenario, is still to be clearly answered [6]. Most data centers typically operate with server inlet temperatures ranging between 18 °C and 24 °C, but we can find some of them as cold as 13 °C [7], and others as hot as 35 °C [8]. These values are often chosen based on conservative suggestions provided by manufacturers, and ensure

* Corresponding author at: DACYA, Universidad Complutense de Madrid, Madrid 28040, Spain.
E-mail addresses: marina.zapater@ucm.es (M. Zapater), jlrisco@ucm.es (J.L. Risco-Martín), parroba@die.upm.es (P. Arroba), jayala@ucm.es (J.L. Ayala), josem@die.upm.es (J.M. Moya), rhermida@ucm.es (R. Hermida).

http://dx.doi.org/10.1016/j.asoc.2016.07.042
1568-4946/

inlet temperatures within the ranges published by ASHRAE (i.e., 15 °C to 32 °C for enterprise servers [9]).

Data center designers have collided with the lack of accurate models for the energy-efficient real-time management of computing facilities. One modeling barrier in these scenarios is the high number of variables potentially correlated with temperature that prevent the development of macroscopic analytical models. Nowadays, to simulate the inlet temperature of servers under certain cooling conditions, designers rely on time consuming and very expensive Computational Fluid Dynamics (CFD) simulations. These techniques use numerical methods to solve the differential equations that drive the thermal dynamics of the data room. They need to consider a comprehensive number of parameters both from the server and the data room (i.e. specific characteristics of servers such as airflow rates, data room dimensions and setup). Moreover, they are not robust to changes in the data center (i.e. rack placement and layout changes, server turn-off, inclusion of new servers, etc.). If the simulation fails to properly incorporate a relevant parameter, or if there is a deviation between the theoretical and the real values, the simulation becomes inaccurate. Due to the high economic and computational cost of CFD simulation, models cannot be re-run each time there is a change in the data room.

To minimize cooling costs, the development of models that accurately predict the CPU temperature of the servers under variable environmental conditions is a major challenge. These models need to work on runtime, adapting to the changing conditions of the data room automatically re-training if data center conditions change dramatically, and enabling data center operators to increase room temperature safely.

The nature of the problem suggests the usage of meta-heuristics instead of analytical solutions. Meta-heuristics make few assumptions about the problem, providing good solutions even when they have fragmented information. Some meta-heuristics such as Genetic Programming (GP) perform Feature Engineering (FE), a particularly useful technique to select the set of features and combination of variables that best describe a model. Grammatical Evolution (GE) is an evolutionary computation technique based on GP used to perform symbolic regression [10]. This technique is particularly useful to provide solutions that include non-linear terms offering Feature Engineering capabilities and removing analytical modeling barriers. Also, designer's expertise is not required to process a high volume of data as GE is an automatic method. However, GE provides a vast space of solutions that may need to be bounded to achieve algorithm efficiency.

This paper develops a data center room thermal modeling methodology based on GE to predict on runtime, and with sufficient anticipation, the critical variables that drive reliability and cooling power consumption in data centers. Particularly, the main contributions of our work are the following:

- The development of multi-variable models that incorporate time dependence based on Grammatical Evolution to predict CPU and inlet temperature of the servers in a data room during runtime. Due to the feature engineering and symbolic regression performed by GE, our models incorporate the optimum selection of representative features that best describe the thermal behavior.
- We prevent premature convergence by means of Social Disaster Techniques and Random Off-Spring Generation, dramatically reducing the number of generations needed to obtain accurate solutions. We tune the models by selecting the optimum parameters and fitness function using a reduced experimental setup, consisting of real measurements taken from a single server isolated in a fully sensorized data room.
- We offer a comparison with other techniques commonly used in literature to solve temperature modeling problems, such as autoregressive moving average (ARMA) models, linear model identification methods (N4SID), and dynamic neural networks (NARX).
- The proposal of an automatic data room thermal modeling methodology that scales our solution to a realistic Data Center scenario. As a case study, we model CPU and inlet temperatures using real traces from a production data center.

Our work allows the generation of accurate temperature models able to work on runtime and adapt to the ever changing conditions of these scenarios, while achieving very low average errors of 2 °C for CPU temperature and 0.5 °C for inlet temperature.

The remainder of the paper is organized as follows: Section 2 accurately describes the modeling problem, whereas Section 3 provides an overview of the current solutions. Section 4 describes our proposed solution, whereas Section 5 presents the experimental methodology. Section 6 shows the results obtained and Section 7 discusses them. Finally, Section 8 concludes the paper.

2. Problem description

2.1. Data room thermal dynamics

To ensure the safe operation of a traditional raised-floor air-cooled Data Center, data rooms are equipped with chilled-water Computer Room Air Conditioning (CRAC) units that use conventional air-cooling methods. Servers are mounted in racks on a raised floor. Racks are arranged in alternating cold/hot aisles, with server inlets facing cold air and outlets creating hot aisles. CRAC units supply air at a certain temperature and air flow rate to the Data Center through the floor plenum. The floor has some perforated tiles through which the blown air comes out. Cold air refrigerates servers and heated exhaust air is returned to the CRAC units via the ceiling, as shown in Fig. 1.

Even though this solution is very inefficient in terms of energy consumption, the majority of the data centers use this mechanism. In fact, despite the recent advances in high-density cooling techniques, according to a survey by the Uptime Institute, in 2012 only 19% of large scale data centers had incorporated other cooling mechanisms [5]. In some scenarios, the control knob of the cooling subsystem is the cold air supply temperature, whereas in others, it is the return temperature of the heated exhaust air to the CRAC unit.

The maximum IT power density that can be deployed in the Data Center is limited by the perforated tile airflow. Because the plenum is usually obstructed (e.g. blocked with cables in some areas), a non-uniform airflow distribution is generated and each tile exhibits a different pressure drop. Moreover, in data centers where the hot and cold aisles are not isolated, which is the most common scenario, the heated exhaust air recirculates to the cold aisle, mixing with the cold air.

2.2. Temperature-energy tradeoffs

The factor limiting minimum data room cooling is maximum server CPU temperature. Temperatures higher than 85 °C can cause permanent reliability failures [11]. At temperatures above 95 °C, servers usually turn off to prevent thermal redlining. Previous work on server power and thermal modeling [12], shows how CPU temperature is dominated by: (i) power consumption, which is dependent on workload execution, (ii) fan speed, which changes the cooling capacity of the server, and (iii) server cold air supply (inlet temperature).

Thus, to keep all the equipment under normal operation, CRAC units have to supply the air at an adequately low temperature to ensure that all CPU's are below the critical threshold. However, inlet