Control Engineering Practice 107 (2021) 104674

Contents lists available at ScienceDirect

Control Engineering Practice

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

Adaptive Predictive Control of a data center cooling unit

Fernando Martínez-García [a,*,1], Ghada Badawy [b], Masoud Kheradmandi [a,1], Douglas G. Down [a,1]

[a] Department of Computing and Software, McMaster University, Hamilton, ON L8S 4L8, Canada
[b] Computing Infrastructure Research Centre, McMaster University, Hamilton, ON L8P 0A1, Canada

ARTICLE INFO

Keywords:
Adaptive Predictive Controller
Variable prediction horizon
Variable forgetting factor
Data center
Temperature control

ABSTRACT

As server densities increase to support the rising demand for high density computing, traditional room cooling infrastructure is struggling to keep up. Rack-mounted cooling units are increasingly being deployed in data centers. These cooling units are located closer to the heat sources (the servers) which allow them to cool more efficiently than traditional cooling infrastructure. In this work, a Multi-Input Single-Output (MISO) Adaptive Predictive Controller (APC) for rack-mounted cooling units is investigated and implemented using a low-cost general-purpose microcontroller. The proposed APC is implemented using Weighted Recursive Least Squares (WRLS) and a sub-optimal but fast algorithm based on the General Predictive Controller (GPC) approach. These are combined with a variable forgetting factor and variable prediction horizon algorithms. In addition, methods are proposed to handle stability issues arising due to practical hardware limitations. The controller is implemented on a real single rack system and challenges for practical implementation are addressed and illustrated. The proposed APC is also compared (via simulation) to a standard APC with equivalent complexity and a split-range PI controller. The results show that the proposed controller outperforms both the standard APC and the split-range PI controllers with respect to Mean Squared Error (MSE).

1. Introduction

Data center (DC) energy consumption has attracted a lot of attention in recent years. According to Koomey (2011), DC energy consumption ranges from 1.1% to 1.5% of total global electricity consumption, with this proportion showing a tendency to increase (Whitney & Delforge, 2014). A significant portion of this energy utilization is devoted to cooling systems that aim to keep server temperatures within a safe region, necessary to avoid damage to servers. The safe temperature ranges for different classes of equipment are provided by guidelines such as those of ASHRAE (2011). Traditionally in DCs, cooling infrastructure is either room-based or row-based (Dunlap & Rasmussen, 2012; Patterson, 2008). However, in recent years rack mountable cooling units have been introduced to cope with the increasing demand for high performance computing (HPC). These new cooling units bring servers and cooling units (Dunlap & Rasmussen, 2012) closer to each other with an aim to decrease cooling infrastructure energy consumption.

In addition to reducing energy consumption, maintaining a stable temperature inside a data center is crucial since oscillations in air temperature, even by 1 or 2 degrees, increase the probability of server failures (El-Sayed, Stefanovici, Amvrosiadis, Hwang, & Schroeder, 2012). These oscillations are an inherent characteristic of the ON/OFF or PID controllers which have been widely used in cooling infrastructure (Afram & Janabi-Sharifi, 2014). Due to the proximity of the rack-mounted cooling units to the servers, any variation in airflow created by these controllers, as a response to changes in workload, will be experienced immediately by the servers, which will consequently lead to higher server failure rates.

Model Predictive Control (MPC) has the potential to avoid these oscillations by estimating the behavior of the system. MPC approaches with a fixed model have shown good results. For instance, the implementation of this type of MPC in DC problems has already been explored (Fang, Wang, & Gong, 2016), showing good results via simulations. Although MPC can handle a wide variety of problems, it usually requires the solution of an iterative optimization problem at each sampling instant, which might make it infeasible in real world deployments since it can require significant computational power (both in terms of speed and memory). This is in contrast with non-iterative controllers, that can handle this computational effort in real-time. Moreover, mismatches due to fixed models in MPC could have negative effects on the performance of the system. Model Predictive Control with re-identification has also been proposed (Kheradmandi & Mhaskar, 2018), where past information is used to recompute the parameters of the system model based on the accuracy of previous predictions of the system behavior, showing good results in simulation. However, the complexity of computations (the required iterative optimization and computation of pseudoinverses of several matrices) as well as the need to store observability and controllability matrices, make it also computationally expensive for real-time implementation. Therefore,

* Corresponding author.
E-mail addresses: martinef@mcmaster.ca (F. Martínez-García), badawyg@mcmaster.ca (G. Badawy), downd@mcmaster.ca (D.G. Down).

[1] This work is a collaborative effort between the Computing Infrastructure Research Centre and the Department of Computing and Software at McMaster University.

https://doi.org/10.1016/j.conengprac.2020.104674
Received 20 December 2019; Received in revised form 29 October 2020; Accepted 31 October 2020
Available online xxxx