University of Groningen

Optimized Thermal-Aware Job Scheduling and Control of Data Centers

Van Damme, Tobias; De Persis, Claudio; Tesi, Pietro

Published in:
IEEE Transactions on Control Systems Technology

DOI:
10.1109/TCST.2017.2783366

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Final author's version (accepted by publisher, after peer review)

Publication date:
2019

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Van Damme, T., De Persis, C., & Tesi, P. (2019). Optimized Thermal-Aware Job Scheduling and Control of Data Centers. IEEE Transactions on Control Systems Technology, 27(2), 760-771. https://doi.org/10.1109/TCST.2017.2783366

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 29-04-2019

Optimized Thermal-Aware Job Scheduling and Control of Data Centers

Tobias Van Damme, Claudio De Persis and Pietro Tesi

***Abstract*—Analyzing data centers with thermal-aware optimization techniques is a viable approach to reduce energy consumption of data centers. By taking into account thermal consequences of job placements among the servers of a data center, it is possible to reduce the amount of cooling necessary to keep the servers below a given safe temperature threshold. We set up an optimization problem to analyze and characterize the optimal setpoints for the workload distribution and the supply temperature of the cooling equipment. Furthermore under mild assumptions we design and analyze controllers that regulate the system to the optimal state without knowledge of the current total workload to be handled by the data center. The response of our controller is validated by simulations and convergence to the optimal setpoints is achieved under varying workload conditions.**

***Index Terms*—Cyber-physical systems, Lyapunov methods, Optimization and control of Large-scale networked systems, Optimization and control of data centers, Control of constrained systems, Networked control systems**

I. Introduction

Worldwide energy consumption of data centers reached 350 billion kWh of energy in 2013, 1.73% of the global electricity consumption [2], [3]. With the world being digitized more and more each year, this number is likely to increase as well. Therefore in the last decade computer scientists and control engineers have made efforts to reduce the energy consumption of data centers by devising methods to increase the operational efficiency of these computer halls [4].

Although much progress has been made, there are still several challenges in ensuring efficient operation of the cooling equipment. Due to bad design or unawareness for the thermal properties of the data center, local thermal hotspots can arise. This causes the cooling equipment to overreact to ensure that the temperature of the equipment stays below the safe thermal threshold. These peaks cause the cooling equipment to consume more energy then would be necessary if these hotspots were avoided. Therefore having a good understanding of the thermodynamics involved is vital to increasing the cooling efficiency of the data center.

To tackle these challenges researchers and engineers have studied both software and hardware solutions to this problem. Examples of hardware solutions are isolating cold or hot areas in the data center, or building data centers in cold regions on the planet where cold outside air can be utilized. Software solutions on the other hand focus on strategies which use the knowledge of the thermal properties of the data center to make more intelligent choices how to schedule incoming jobs. Although the two types of solutions are equally important to study, software solutions allow data center operators to implement improvements very fast for very little costs, i.e. implementing new software is less costly than rebuilding a full data center. Furthermore these types of solutions can provide major performance increases using heuristic methods for smart thermal job schedulers showing up to 30% less energy consumption with respect to non thermal-aware job schedulers [5], [6]. Therefore in this paper we will investigate these thermal-aware software solutions.

Other approaches include considering a heterogeneous data center [7] and using these asymmetric properties to analyze trade-offs between performance- and energy-aware algorithms, or distinguishing between different type of jobs when scheduling the load [8]. Server consolidation is a natural extension where on top of thermal scheduling, racks are switched on and off to save power. These algorithms usually contain two steps, first to calculate the necessary number of racks and secondly the correct workload scheduling [5], [9], [10], [11], [12].

On the other hand, studies have also been done in a more control theoretical direction. The paper [13] has proposed a control algorithm that tries to maintain the temperature of the equipment around a target value. In [14] it is proposed to implement a two-step algorithm that first minimizes the energy consumption by estimating the required amount of servers to handle the expected workload. In the second step the algorithm maximizes the response time given a number of servers at its disposal. In an attempt to address scalability a distributed approach has been studied in [15]. In this work, units, which range from servers to complete data centers, communicate directly and try to achieve a uniform temperature profile. Another distributed control approach in a hybrid systems setting is proposed in [16]. The hybrid controller tries to evenly divide the total load among the agents in the network in a distributed fashion.

While all these works have strong points on their own, to the authors' best knowledge a thorough formal analysis and characterization of an energy minimal solution combined with a control strategy which handles both cooling and job scheduling simultaneously has not been done before. The objective of this work is to provide an extendable framework that allows for a characterization of an energy-minimal operating point and then supply methods for operating the data center such that this operating point is achieved for all load conditions.

The contribution of this work is two-fold. First from existing thermodynamical principles we set up a thermodynamical model from which we derive an optimization problem that

All authors are with the Department of ENTEG, Faculty of Mathematics and Natural Sciences, University of Groningen, 9747 AG Groningen, The Netherlands. {t.van.damme, c.de.persis, p.tesi}@rug.nl

The authors declare no competing financial interest

An abridged version of this paper has appeared in [1].

combines energy minimization with the thermodynamics. In addition to only including temperature constraints [11] we extend the model to also incorporate workload constraints, which allows us to better characterize energy minimal solutions. This design allows for natural extendability to more complicated scheduling policies like switching servers on and off.

Second we develop a novel control strategy for handling the control of the cooling equipment and the workload scheduling simultaneously. Both these control goals have been studied before [13], [17]. However in [13] the two control goals were handled separately; in [17] a combined algorithm was suggested but due to complexity could lead to non-optimal solution. In contrast our model shows an easy method for handling coordinated cooling and job scheduling control that regulates the system towards the energy optimal solution. Our method is inspired by results from [18] where regulation to optimal steady solutions in the presence of disturbances was considered.

The remainder of this paper is organized as follows. In Section II the basic thermodynamics are formulated. Then an optimization problem is formulated in Section III and under mild assumptions the equivalence to a reduced form is proven. Following up the optimal solution is analytically analyzed and characterized for different load conditions in Section IV. Using this analytical solution a controller is proposed in Section V that can handle unknown load conditions. Finally in Section VI a case study is considered to show the performance of the controllers.

Notation: We denote by $\mathbb{R}$ and $\mathbb{R}_{\geq 0}$ the set of real numbers and non-negative real numbers respectively. Vectors and matrices are denoted by $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times m}$ respectively, the transpose is denoted by x^T and the inverse of a matrix is denoted by A^{-1}. If the entries of x are functions of time then the element-wise time derivative is denoted by $\dot{x}(t) := \frac{d}{dt}x$. By x_i we denote the i-th element of x and by a_{ij} we denote the ij-th element of A. If a variable already has another subscript then we switch to superscripts to denote individual elements, i.e. T^i_{out} and C^{ij}_3. We write the diagonal matrix constructed from the elements of vector x as $\text{diag}\{x_1, \quad x_2, \quad \cdots, \quad x_n\}$. The identity matrix of dimension n is denoted by I_n, the vector of all ones by $\mathbb{1} \in \mathbb{R}^n$ and the vector of all zeros by $\mathbb{0} \in \mathbb{R}^n$. Furthermore the vector comparison $x \preccurlyeq y$ is defined as the element-wise comparison $x_i \leq y_i$ for all elements in x and y.

II. System model

Real life data centers are organized in aisles with many racks each containing a multitude of servers. The cooling of data centers is usually done by air conditioning, therefore the racks are set up in a hot- and cold-aisle configuration. Cold air supplied by the computer room air conditioning (CRAC) units is blown into the cold aisles. The air goes through the racks where it absorbs the heat produced by the servers. The air exits the servers in the hot aisle and is recirculated back to the CRAC units where it is cooled down to the desired supply temperature. A scheduler divides incoming tasks among the racks according to some decision policy. The energy consumption of a rack depends on the amount of tasks it is given. By thermodynamical principles almost all of this energy consumption is dissipated as heat in the rack. Ideally all of the exhaust air of the racks is returned to the CRAC, however due to the complex nature of air flows within the data center some of the hot air is recirculated back into the cold aisles. This causes the temperature of the air at the inlet of the racks to rise, creating inefficiencies in the cooling of the data center.

It is possible for data centers to have multiple CRAC units. In these cases we assume that the CRAC units work as one. Allowing different setpoints will result in mixing of air flows of different temperatures. Air flows of different temperatures however are highly non-linear flows which depend on the temperature of the flow itself. Therefore allowing different CRAC setpoints makes air flows difficult to model and thus adds increased complexity to the already complex situation. This added complexity goes beyond the scope of the paper and as such we do not pursue this possibility.

A. Workload

Requests arriving at the data center are collected by a scheduler which then decides according to some policy how to divide this work among the available racks. We assume that each job has an accompanying tag which denotes the time and the number of computing units (CPU) it requires for execution. Let J denote the integer number of jobs that the scheduler has to schedule in the data center at time t. Then $\mathcal{J}(t) = \{1, \cdots, J\}$ denotes the set of jobs to be scheduled at time t. Furthermore let λ_j be the number of CPU's that job j requires at time t. Then the total number of CPU's, D^*, the scheduler has to divide over the racks at time t is given by

$$D^*(t) = \sum_{j=1}^{\mathcal{J}(t)} \lambda_j. \tag{1}$$

We denote by $D_i(t)$ the number of CPU's the schedulers assigns to rack i at time t. These variables are collected in the vector

$$D(t) := \begin{pmatrix} D_1(t) & D_2(t) & \cdots & D_n(t) \end{pmatrix}^T.$$

B. Power consumption of racks

Different ways to model power consumption exist, with the main difference being the scope and focus of the models. There are models which try to go as close to the CPU level as possible by modeling the power consumption as a function of the CPU clock frequency. On the other hand, there are other models that aim at modeling the system on a higher level and capture the power consumption of the CPU as a function of the workload imposed on the server. The models trade between complexity and detail, where the CPU frequency model captures more details, but results in a non-linear model, and the workload model results in a linear model which operates on a higher level, i.e. the server level. Because in this paper we work at the rack level, a higher operating level in a data center environment, the linear model fits much better to

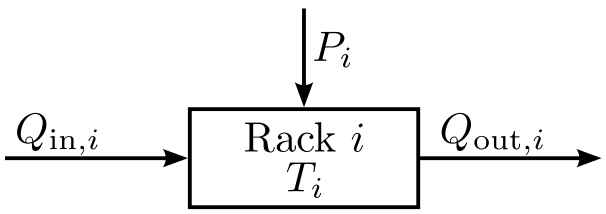

Fig. 1. Heat model of an individual rack. Q^i_{in} is the heat entering the rack, Q^i_{out} is the heat exiting the rack and P_i is the power consumption of the rack.

our situation. This model has been studied many times before and the accuracy loss is almost negligible, as it has been found that these models are about 95% accurate [11], [19], [20], [21], [22], [23], [24], [25]. Therefore this is our model of choice.

Let $P_i(t)$ denote the power consumption of rack i at time t. We model $P_i(t)$ to consist of a load-independent part, e.g. the equipment consumes a constant amount of power, and a load-dependent part, e.g. the number of CPU's that are actively processing jobs

$$P_i(t) = v_i + w_i D_i(t), \tag{2}$$

where v_i [Watts] is the power consumption for the unit being powered on, w_i [Watts CPU^{-1}] is the power consumption per CPU in use. The variables are collected in the vectors

$$P(t) := \begin{pmatrix} P_1(t) & P_2(t) & \cdots & P_n(t) \end{pmatrix}^T,$$
$$V := \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix}^T,$$

and

$$W := \text{diag}\{w_1, \quad w_2, \quad \cdots, \quad w_n\},$$

so that

$$P(t) = V + WD(t). \tag{3}$$

C. Thermodynamical model

Following similar arguments as in [13] and [23], a thermodynamical model for each individual rack is constructed. For our model we focus on the output temperature of the racks as we study the thermodynamical coupling between the workload that is processed by the servers and the energy efficiency of the cooling equipment. As we will show below there is a direct coupling between the output temperature of the racks and both these elements. In Fig. 1 a graphical representation of the heat flows involved is given. The change of temperature of a rack is given by the difference in heat entering and exiting the rack,

$$m_i c_p \frac{d}{dt} T^i_{\text{out}}(t) = Q^i_{\text{in}}(t) - Q^i_{\text{out}}(t) + P_i(t). \tag{4}$$

Here T^i_{out} [°C] is the temperature of the exhaust air at rack i, c_p [J °C^{-1} kg^{-1}] is the specific heat capacity of air, m_i [kg] is the mass of the air inside the rack, Q^i_{in} [Watts] and Q^i_{out} [Watts] are the heat entering and exiting the rack respectively. The heat that enters a rack consists of two parts due to the complex air flows in the data center, i.e. the recirculated air originating from the other racks and the cooled air supplied by the CRAC

$$Q^i_{\text{in}}(t) = \sum_{j=1}^n \gamma_{ji} Q^j_{\text{out}}(t) + Q^i_{\text{sup}}(t). \tag{5}$$

Here Q^i_{sup} [Watts] is the heat supplied by the CRAC to rack i, and γ_{ji} is the percentage of the flow which recirculates from rack j to rack i. The relation between heat and temperature is given by

$$Q(t) = \rho c_p f T(t), \tag{6}$$

where ρ [kg m^{-3}] is the density of the air and f [m^3 s^{-1}] is the flow rate of the given flow. Combining (5) and (6) with (4) yields

$$\frac{d}{dt} T^i_{\text{out}}(t) = \frac{\rho}{m_i} \left(\sum_{j=1}^n \gamma_{ji} f_j T^j_{\text{out}}(t) - f_i T^i_{\text{out}}(t) \right) + \frac{\rho}{m_i} \left(f_i - \sum_{j=1}^n \gamma_{ji} f_j \right) T_{\text{sup}}(t) + \frac{1}{m_i c_p} P_i(t), \tag{7}$$

where T_{sup} [°C] is the temperature of the air supplied by the CRAC and f_i is the velocity of the air flow through rack i. Rewriting the above relation in matrix form, i.e. combining the temperature changes of all racks in one equation, results in

$$\frac{d}{dt} T_{\text{out}}(t) = A(T_{\text{out}}(t) - \mathbb{1} T_{\text{sup}}(t)) + M^{-1} P(t). \tag{8}$$

Here

$$T_{\text{out}}(t) := \begin{pmatrix} T^1_{\text{out}}(t) & T^2_{\text{out}}(t) & \cdots & T^n_{\text{out}}(t) \end{pmatrix}^T,$$

and

$$A := \rho c_p M^{-1} (\Gamma^T - I_n) F,$$
$$F := \text{diag}\{f_1, \quad f_2, \quad \cdots, \quad f_n\},$$
$$M := \text{diag}\{c_p m_1, \quad c_p m_2, \quad \cdots, \quad c_p m_n\},$$
$$\Gamma := [\gamma_{ij}]_{n \times n}.$$

D. Power consumption of CRAC

The power consumption of the CRAC is dependent on the temperature of the air which is returned to CRAC and the supply temperature it has to provide. The air flow which is returned from rack i to the CRAC is given by

$$f^{\text{ret}}_{\text{sup},i} = \left(1 - \sum_{j=1}^n \gamma_{ij} \right) f_i, \tag{9}$$

and therefore the heat returned from all the racks to the CRAC is

$$Q_{\text{ret}}(t) = \rho c_p \sum_{i=1}^n \left(1 - \sum_{j=1}^n \gamma_{ij} \right) f_i T^i_{\text{out}}(t). \tag{10}$$

The heat the CRAC sends back to the data center is given by $Q_{\text{sup}}(t) = \rho c_p f_{\text{sup}} T_{\text{sup}}(t)$. With this the heat the CRAC has to remove from the air, $Q_{\text{rem}}(t)$, is given by

$$\begin{aligned} Q_{\text{rem}}(t) &= Q_{\text{ret}}(t) - Q_{\text{sup}}(t) \\ &= \rho c_p \sum_{i=1}^n \left[\left(1 - \sum_{j=1}^n \gamma_{ij} \right) f_i (T^i_{\text{out}}(t) - T_{\text{sup}}(t)) \right] \\ &= -\mathbb{1}^T M A (T_{\text{out}}(t) - \mathbb{1} T_{\text{sup}}(t)). \end{aligned} \tag{11}$$

To determine the amount of work the CRAC has to do to remove a certain amount of heat, Moore et al. [5] introduced the Coefficient of Performance, $\text{COP}(T_{\text{sup}}(t))$, to indicate the efficiency of the CRAC as a function of the target supply temperature. They found that CRAC units work more efficiently when the target supply temperature is higher. The COP represents the ratio of heat removed to the amount of work necessary to remove that heat. For a water-chilled CRAC unit in the HP Utility Data Center they found that the COP is a quadratic, increasing function. In a general sense the COP can be any monotonically increasing function. The power consumption of the CRAC units can then be given by

$$P_{AC}(T_{\text{out}}(t), T_{\text{sup}}(t)) = \frac{Q_{\text{rem}}(t)}{\text{COP}(T_{\text{sup}}(t))}. \tag{12}$$

Assumption 1. The function $\text{COP}(T_{\text{sup}})$ of the CRAC unit considered in this paper, is monotonically increasing in the range of operation for T_{sup}. □

Example 1. Let us consider a small example to illustrate the influence of a small difference in supply temperature on the power consumption of the CRAC. Consider the quadratic $\text{COP}(T_{\text{sup}}(t))$ found by [5], and two cases where the returned air has to be cooled down by 5 °C, in the first case from 25 °C to 20 °C and in the second case from 30°C to 25°C. Assume that the energy contained in 5 °C temperature difference of air is 100 Watts. In the first case $\text{COP}(20) = 3.19$ and in the second case $\text{COP}(25) = 4.73$. By (12), the energy consumed by the CRAC to cool down the returned air to the required temperature is

$$P_{AC,1} = \frac{100}{3.19} = 31.34 \text{ W}, \quad P_{AC,2} = \frac{100}{4.73} = 21.14 \text{ W}.$$

Here it seen that if the temperature of the returned air increases by 5 °C the power consumption of the CRAC unit decreases by 30%. □

Having completed the model finally allows us to formulate the control problem we would like to solve.

III. Problem formulation

The objective of this paper is two-fold, first we want to find optimal setpoints for the temperature distribution, the supply temperature and workload distribution that minimize the power consumption of the data center. Secondly we want to design controllers which ensure convergence of the variables to the obtained setpoints. Hence the control problem is defined as follows:

Problem 1. For system (8) design controllers for the workload distribution $D(t)$ and supply temperature $T_{\text{sup}}(t)$ such that, given an unmeasured total load $D^*(t)$, any solution of the closed-loop system is bounded and satisfies

$$\lim_{t\to\infty} (T_{\text{out}}(t) - \bar{T}_{\text{out}}) = 0, \tag{13}$$

$$\lim_{t\to\infty} (T_{\text{sup}}(t) - \bar{T}_{\text{sup}}) = 0, \tag{14}$$

$$\lim_{t\to\infty} (D(t) - \bar{D}) = 0, \tag{15}$$

where $\bar{T}_{\text{out}}$, $\bar{T}_{\text{sup}}$ and $\bar{D}$ are the optimal setpoint values for the temperature distribution, supply temperature and the power consumption, i.e. workload distribution, respectively, which are defined in Subsection III-A. □

From this point on we will implicitly assume the dependence of the variables on time and only denote it when confusion might arise otherwise.

A. General optimization problem

We formulate an optimization problem to minimize the power consumption while taking into account the physical constraints of the equipment, i.e the servers only have finite computational capacity and the temperature of the servers cannot exceed a certain threshold. The power consumption of the data center can be written as a combination of 2 parts, the power consumption of the cooling equipment and the power consumption of the racks. Combining (3) and (12) we can write the total power consumption as

$$\mathcal{C}(T_{\text{out}}, T_{\text{sup}}, D) = \frac{Q_{\text{rem}}}{\text{COP}(T_{\text{sup}})} + \mathbb{1}^T P(D). \tag{16}$$

A reasonable way [11], [14] to formulate the optimization problem is

$$\min_{T_{\text{out}}, T_{\text{sup}}, D} \quad \frac{Q_{\text{rem}}}{\text{COP}(T_{\text{sup}})} + \mathbb{1}^T P(D) \tag{17a}$$

$$s.t. \quad D^* = \mathbb{1}^T D \tag{17b}$$

$$\mathbb{0} \preccurlyeq D \preccurlyeq D_{\text{max}} \tag{17c}$$

$$\mathbb{0} = A(T_{\text{out}} - \mathbb{1}T_{\text{sup}}) + M^{-1}P(D) \tag{17d}$$

$$T_{\text{out}} \preccurlyeq T_{\text{safe}}. \tag{17e}$$

Equation (17b) ensures that all the available work is divided among the racks, (17c) encompasses the computational capacity of the rack, i.e. rack i has D^i_{max} CPU's available at most. The system dynamics should be at steady state once the optimal point has been reached, see (17d), and finally (17e) enforces that the temperature of the racks is below the given safe threshold, $T_{\text{safe}} \in \mathbb{R}^n$.

B. Equivalent optimization problem for homogeneous data centers

Due to the non-linear nature of how the COP affects the power consumption it is not trivial to analyze the general optimization problem. Although (17) is a difficult problem to solve analytically, it is possible to reduce the optimization problem to a simpler equivalent problem for a specific important case. In many of the larger real-life data centers most of the equipment is identical, i.e. the power consumption characteristics of the computational equipment is identical, that is $v_i = v$ and $w_i = w$ for all i in (2). It is desirable for data centers to employ identical equipment because this allows for decreased maintenance complexity and allows for bulk purchases of the equipment which reduce operational costs. In this case the data center is said to be composed of homogeneous racks or, more simply, the data center is homogeneous.

In case of a homogeneous data center the power consumption is given by $P(D) = v\mathbb{1} + wD$ and the total computational power consumption is given by

$$\mathbb{1}^T P(D) = nv + w\mathbb{1}^T D = nv + wD^*. \tag{18}$$

For this case, the computational power consumption no longer depends on the way the jobs are distributed but only depends on the total workload. This property simplifies the cost function defined in (16) considerably.

Theorem 1. *Let the data center consist of homogeneous racks, i.e. $v_i = v$, and $w_i = w$ for all i in* (2) *and assume constraint* (17d) *is satisfied. Then problem* (17) *is equivalent to*

$$\max_{T_{out}} \quad C_1^T T_{out} \tag{19a}$$
$$s.t. \quad \mathbb{0} \preccurlyeq C_3 T_{out} + C_4(D^*) \preccurlyeq D_{max} \tag{19b}$$
$$T_{out} \preccurlyeq T_{safe}, \tag{19c}$$

for suitable C_1, C_3 and C_4. □

Before we prove this theorem we need to introduce some notation and extra preparatory results. In these preparatory results (Lemma 1-3 below), the homogeneity condition is not required, and statements are given in terms of the power consumption vector P defined as in (3).

Lemma 1. *Equation* (17d) *implies that the following relation holds*

$$\mathbb{1}^T P(D) = -\mathbb{1}^T MA(T_{out} - \mathbb{1}T_{sup}) = Q_{rem},$$

with Q_{rem} defined in (11)*, which reduces the cost function to*

$$\mathcal{C}(T_{out}, T_{sup}, D) = \left(1 + \frac{1}{COP(T_{sup})}\right)\mathbb{1}^T P(D). \tag{20}$$

Proof. By multiplying (17d) by $\mathbb{1}^T M$ and solving for $\mathbb{1}^T P(D)$ we obtain above result. □

Lemma 2. *If* (17b) *and* (17d) *are satisfied, then*

$$T_{sup} = C_1^T T_{out} + C_2(D^*), \tag{21}$$
$$C_1^T \triangleq: \frac{\mathbb{1}^T W^{-1} MA}{\mathbb{1}^T W^{-1} MA\mathbb{1}},$$
$$C_2(D^*) \triangleq: \frac{D^* + \mathbb{1}^T W^{-1} V}{\mathbb{1}^T W^{-1} MA\mathbb{1}}.$$

Proof. After multiplying (17d) by $\mathbb{1}^T W^{-1} M$, combining with (17b) and some basic matrix manipulations the result is obtained. □

Lemma 3. *If* (17b) *and* (17d) *are satisfied, then*

$$D = C_3 T_{out} + C_4(D^*), \tag{22}$$
$$C_3 \triangleq: -W^{-1} MA(I_n - \mathbb{1}C_1^T),$$
$$C_4(D^*) \triangleq: W^{-1} MA\mathbb{1}C_2(D^*) - W^{-1} V.$$

Proof. Substituting the result of Lemma 2 in (17d), pre-multiplying (17d) by $W^{-1}M$, and solving for D yields the result. □

Remark 1. The dimensions of the constants from above Lemmas are $C_1 \in \mathbb{R}^n$, $C_2 \in \mathbb{R}$, $C_3 \in \mathbb{R}^{n\times n}$ and $C_4 \in \mathbb{R}^n$. The following identities for the constants C_1, C_3 and C_4 are observed

$$C_1^T \mathbb{1} = 1, \quad \mathbb{1}^T C_3 = \mathbb{0}^T, \quad C_3\mathbb{1} = \mathbb{0}, \quad \mathbb{1}^T C_4 = D^*. \tag{23}$$

An important consequence worth to note is that the constant $\mathbb{1}^T D$, with D defined as in (22), satisfies the identity $\mathbb{1}^T D = D^*$. □

Lemma 2 and Lemma 3 show that at the steady state the supply temperature, T_{sup}, and workload distribution vector, D, are uniquely defined by the total workload, D^*, and the temperature distribution, T_{out}. With these properties in mind we are ready to prove Theorem 1.

Proof of Theorem 1. Assume that Problem (17) has a solution. By Lemma 1, the cost function reduces to (20). By the homogeneity assumption, (18) holds, which shows that the cost function (20) is independent of the distribution D and depends only on T_{sup}. Hence, in view of Assumption 1 (monotonicity of the function COP(T_{sup})) a solution to Problem (17) is the one that maximizes T_{sup}. By (21) in Lemma 2, this solution must maximize the cost function in (19a). The constraints in (17) and Lemma 3 imply the constraints in (19), showing that a solution to (17) must be also a solution to (19).

Conversely, if a solution to (19) exists, define D as in (22), and notice that (17b) is satisfied, as it is promptly verified using the identities (23). Then by the homogeneity assumption, (17d), Lemma 1, and Lemma 2, maximizing the cost function in (19a) implies minimizing the cost function in (17a). Moreover, the definition of D and the constraint (19b) implies (17c). Constraint (17e) trivially holds because of (19c). This ends the proof. □

IV. Characterization of the optimal solution

In the previous section we have showed the possibility to reduce the optimization problem to a simpler form. In this section we show that using KKT optimality conditions it is possible to further characterize the optimal point.

A. KKT optimality conditions

Because the optimization problem (19) is convex and all inequality constraints are linear functions we have that Slater's condition holds. Therefore it follows that $\bar{T}_{\text{out}}$ is an optimal solution to (19) if and only if there exists $\bar{\mu}, \bar{\mu}_+, \bar{\mu}_- \in \mathbb{R}^n_{\geq 0}$ such that the following set of relations is satisfied [26]:

$$-C_1 + \bar{\mu} + C_3^T(\bar{\mu}_+ - \bar{\mu}_-) = \mathbb{0}, \tag{24a}$$
$$\mathbb{0} \preccurlyeq C_3\bar{T}_{\text{out}} + C_4(D^*) \preccurlyeq D_{\text{max}}, \tag{24b}$$
$$\bar{T}_{\text{out}} \preccurlyeq T_{\text{safe}}, \tag{24c}$$
$$\bar{\mu}_+^T(C_3\bar{T}_{\text{out}} + C_4(D^*) - D_{\text{max}}) = 0, \tag{24d}$$
$$\bar{\mu}_-^T(-C_3\bar{T}_{\text{out}} - C_4(D^*)) = 0, \tag{24e}$$
$$\bar{\mu}^T(\bar{T}_{\text{out}} - T_{\text{safe}}) = 0, \tag{24f}$$
$$\bar{\mu}, \bar{\mu}_+, \bar{\mu}_- \succcurlyeq \mathbb{0}. \tag{24g}$$

The Lagrangian corresponding to the optimal problem is given by:

$$\begin{aligned}\mathcal{L}(\mu,\mu_+,\mu_-,T_{\text{out}}) = &- C_1^T T_{\text{out}} + \mu^T(T_{\text{out}} - T_{\text{safe}}) \\ &+ \mu_-^T(-C_3 T_{\text{out}} - C_4(D^*)) \\ &+ \mu_+^T(C_3 T_{\text{out}} + C_4(D^*) - D_{\text{max}}).\end{aligned} \tag{25}$$

B. Characterization of optimal temperature profile

By studying the KKT optimality conditions we can characterize the optimal solution in different cases.

- *Inactive workload constraints:* Every rack is processing some work but not all the processors of each rack are utilized:

$$0 < (C_3\bar{T}_{\text{out}} + C_4(D^*))_i < D^i_{\text{max}} \quad \forall i \in \{1,\cdots,n\}.$$

- *Partially active workload constraints:* In k racks all processors are processing jobs. The other $n-k$ racks are processing some work but still have processors available:

$$\begin{aligned}(C_3\bar{T}_{\text{out}} + C_4(D^*))_i &= D^i_{\text{max}} \quad \forall i \in \{1,\cdots,k\}, \\ 0 < (C_3\bar{T}_{\text{out}} + C_4(D^*))_i &< D^i_{\text{max}} \quad \forall i \in \{k+1,\cdots,n\}.\end{aligned}$$

The optimal temperature profile corresponding to these two cases is summarized in the following two theorems.

Theorem 2. *Assume the case that none of the workload constraints are active, i.e.*

$$0 < (C_3\bar{T}_{out} + C_4(D^*))_i < D^i_{max} \quad \forall i \in \{1,\cdots,n\}.$$

The solution to (24) *and the optimal solution for the optimization problem* (19) *is then given by*

$$\bar{\mu}_+ = \bar{\mu}_- = \mathbb{0}, \quad \bar{\mu} = C_1 \succ \mathbb{0}, \quad \bar{T}_{out} = T_{safe}. \tag{26}$$

Proof. Because all the inequality constraints regarding the workload are inactive we have that both $C_3\bar{T}_{\text{out}} + C_4(D^*) - D_{\text{max}} \prec \mathbb{0}$, and $-C_3\bar{T}_{\text{out}} - C_4(D^*) \prec \mathbb{0}$. Then from (24d) and (24e) we have that $\bar{\mu}_+ = \bar{\mu}_- = \mathbb{0}$. From (24a) it follows that $\bar{\mu} = C_1 \succ \mathbb{0}$ such that from (24f) we conclude that $\bar{T}_{\text{out}} = T_{\text{safe}}$. □

Theorem 3. *In the case that a part of the workload constraints are active, i.e.*

$$\begin{aligned}(C_3\bar{T}_{out} + C_4(D^*))_i &= D^i_{max} \quad \forall i \in \{1,\cdots,k\}, \\ 0 < (C_3\bar{T}_{out} + C_4(D^*))_i &< D^i_{max} \quad \forall i \in \{k+1,\cdots,n\},\end{aligned}$$

the solution of (24) *is as follows:*

(i) For the racks at the constraint boundary, $i \in \{1,\cdots,k\}$:

$$\bar{\mu}^i_- = 0, \quad \frac{C^i_1 + \sum_{j=1,j\neq i}^k \bar{\mu}^j_+ \left|C^{ji}_3\right|}{C^{ii}_3} \geq \bar{\mu}^i_+ \geq 0, \tag{27}$$

$$\bar{\mu}^i = C^i_1 + \sum_{j=1,j\neq i}^k \bar{\mu}^j_+ \left|C^{ji}_3\right| - \bar{\mu}^i_+ C^{ii}_3 \geq 0, \tag{28}$$

$$\begin{aligned}\bar{T}^i_{out} = &\frac{D^i_{max} - C^i_4(D^*)}{C^{ii}_3} + \sum_{j=k+1}^n \frac{\left|C^{ij}_3\right|}{C^{ii}_3} T^j_{safe} \\ &+ \sum_{j=1,j\neq i}^k \frac{\left|C^{ij}_3\right|}{C^{ii}_3} \bar{T}^j_{out} \\ &\leq T^i_{safe}.\end{aligned} \tag{29}$$

(ii) For the racks that are not at the constraint boundary, $i \in \{k+1,\cdots,n\}$:

$$\bar{\mu}^i_- = \bar{\mu}^i_+ = 0, \tag{30}$$

$$\bar{\mu}^i = C^i_1 + \sum_{j=1}^k \bar{\mu}^j_+ \left|C^{ji}_3\right| > 0, \tag{31}$$

$$\bar{T}^i_{out} = T^i_{safe}. \tag{32}$$

□

Before we can prove Theorem 3 we need to know more about the structure of C_3.

Property 1. Consider C_3 as defined in Lemma 3. The off-diagonal terms of this matrix are strictly negative and the diagonal terms are strictly positive.

Proof. The proof can be found in Appendix A. □

Proof of Theorem 3. Because part of the workload constraints are at the constraint boundary, the analysis following from the Lagrange multipliers is more involved. First we can say that

$$\begin{aligned}\bar{\mu}^i_- &= 0 \quad \forall i, \\ \bar{\mu}^i_+ &= 0 \quad \forall i \in \{k+1,\cdots,n\}, \\ \bar{\mu}^i_+ &\geq 0 \quad \forall i \in \{1,\cdots,k\}.\end{aligned}$$

Then from (24a)

$$\bar{\mu}^i = C^i_1 - \sum_{j=1}^k \bar{\mu}^j_+ C^{ji}_3. \tag{33}$$

From Property 1 we have that the off-diagonal elements of C_3 are strictly negative. For racks $i \in \{k+1,\cdots,n\}$ we have that the C^{ji}_3 elements in (33) will always be off-diagonal elements. Therefore rewriting (33) gives

$$\bar{\mu}^i = C^i_1 + \sum_{j=1}^k \bar{\mu}^j_+ \left|C^{ji}_3\right| > 0 \quad \forall i \in \{k+1,\cdots,n\}, \tag{34}$$

then from (24f) it holds that

$$\bar{T}^i_{\text{out}} = T^i_{\text{safe}} \quad \forall i \in \{k+1,\cdots,n\}. \tag{35}$$

For racks $i \in \{1, \cdots, k\}$ (33) is given by

$$\bar{\mu}^i = C_1^i + \sum_{j=1, j\neq i}^{k} \bar{\mu}_+^j \left|C_3^{ji}\right| - \bar{\mu}_+^i C_3^{ii} \geq 0. \tag{36}$$

For (36) to hold, it should hold that

$$\frac{C_1^i + \sum_{j=1,j\neq i}^{k} \bar{\mu}_+^j \left|C_3^{ji}\right|}{C_3^{ii}} \geq \bar{\mu}_+^i \quad \forall i \in \{1, \cdots, k\}. \tag{37}$$

As the left hand side of (37) is strictly positive for all $i \in \{1, \cdots, k\}$, it is possible to find feasible $\bar{\mu}_+^i \geq 0$ such that $\bar{\mu}^i \geq 0$ for all i. It can be shown that $\bar{T}_{\text{out}}^i$ for all $i \in \{1, \cdots k\}$ is given as

$$\begin{aligned}\bar{T}_{\text{out}}^i =& \frac{D_{\text{max}}^i - C_4^i(D^*)}{C_3^{ii}} + \sum_{j=k+1}^{n} \frac{\left|C_3^{ij}\right|}{C_3^{ii}} T_{\text{safe}}^j \\ &+ \sum_{j=1,j\neq i}^{k} \frac{\left|C_3^{ij}\right|}{C_3^{ii}} \bar{T}_{\text{out}}^j \\ &\leq T_{\text{safe}}^i. \end{aligned} \tag{38}$$

□

Remark 2. One cannot freely choose the k racks for which $D_i = D_{\text{max}}^i$. Whether or not a rack is processing its maximum capacity depends on the data center parameters, i.e. small amount of recirculated air at the input of the rack and low power consumption of the computational equipment. For these racks it holds that

$$\bar{T}_{\text{out}}^i \leq T_{\text{safe}}^i \quad \forall i \in \{1, \cdots, k\}.$$

V. Temperature based job scheduling control

As established in Section IV it is possible to calculate the optimal solution under the assumption that the total workload at time t, D^* is known. However it might not always be possible to obtain this quantity. For example when jobs arrive in the data center in some cases it might be hard to assess how much resources these jobs need. Consider the case where a virtual machine is requested by a user. Usually a certain amount of resources are allocated to this virtual machine, however the user need not use all the available resources all the time. In those situation it is hard to obtain the real workload. In this section we design a controller that is still able to achieve the control goals defined in (13)-(15) under the assumption that $\mathbb{0} \prec D \prec D_{\text{max}}$. From Theorem 2 we see that in this case the optimal solution is always $\bar{T}_{\text{out}} = T_{\text{safe}}$, independent of the way the jobs are distributed. Since most data centers are designed to have overcapacity usually the computational bounds of the racks will not be reached and this assumption is valid in those setups.

A. Controller design

We will now design the control inputs for the workload distribution, D, and the supply temperature of the CRAC unit, T_{sup} while the total workload D^* is unknown. Furthermore the controllers only have access to the measurement of the output temperature of the air at the outlet of each rack, T_{out}. In other words we design temperature feedback algorithms to dynamically adjust D and T_{sup} such that control objectives (13)-(15) are achieved. The proposed controllers for the supply temperature and the workload distribution are given by

$$\dot{T}_{\text{sup}} = \mathbb{1}^T A^T Z (T_{\text{out}} - T_{\text{safe}}), \tag{39}$$

$$\dot{D} = (\frac{\mathbb{1}\mathbb{1}^T}{n} - I_n) B^T Z (T_{\text{out}} - T_{\text{safe}}), \tag{40}$$

where A is Hurwitz, see Appendix B for the proof of this property. Since A is Hurwitz we can find a positive definite matrix Z such that

$$A^T Z + ZA = -2I_n, \tag{41}$$

and B is

$$B = M^{-1} W,$$

where W is defined Subsection II-B, and A and M are defined in Subsection II-C. The controllers (39) and (40) depend only on the output temperature and the system parameters and will continue to vary until the output temperature reaches the safe temperature, which is in line with the control objectives. Intuitively the workload controller will shift jobs between racks based on the temperature deviation until the data center has reached the optimal state. In the results below we discuss the convergence behavior of the controllers in a time frame where the total workload, D^*, is assumed to be constant.

Theorem 4. *Let the data center consist of homogeneous racks, i.e. $v_i = v$, and $w_i = w$ for all i in* (2)*, and assume D^* is constant and $\mathbb{1}^T D(0) = D^*$. Then the solution of system* (8) *with controllers* (39) *and* (40) *is bounded and converges to the optimal solution of the optimal problem defined in* (17) *and therefore satisfies control objectives* (13)-(15)*.*

Proof. For ease of notation we introduce incremental variables to denote deviations from optimal values

$$\begin{aligned}\tilde{T}_{\text{out}} &= T_{\text{out}} - \bar{T}_{\text{out}}, \\ \tilde{T}_{\text{sup}} &= T_{\text{sup}} - \bar{T}_{\text{sup}}, \\ \tilde{D} &= D - \bar{D},\end{aligned}$$

where $\bar{T}_{\text{out}} = T_{\text{safe}}$, $\bar{T}_{\text{sup}}$ as in (21) and $\bar{D}$ defined as the right-hand side of (22). With these definitions system (8) can be rewritten as

$$\dot{\tilde{T}}_{\text{out}} = A\tilde{T}_{\text{out}} - A\mathbb{1}\tilde{T}_{\text{sup}} + B\tilde{D}, \tag{42}$$

where A and B are as before

$$\begin{aligned}A &= \rho c_p M^{-1}(\Gamma^T - I_n)F, \\ B &= M^{-1}W.\end{aligned}$$

Define the incremental storage functions as

$$\Xi_1(\tilde{T}_{\text{sup}}) = \frac{1}{2}\left\|\tilde{T}_{\text{sup}}\right\|^2, \tag{43}$$

$$\Xi_2(\tilde{D}) = \frac{1}{2}\left\|\tilde{D}\right\|^2. \tag{44}$$

The storage functions satisfy

$$\dot{\Xi}_1(\tilde{T}_{\text{sup}}, \tilde{T}_{\text{out}}) = \tilde{T}_{\text{sup}}^T \dot{T}_{\text{sup}} = \tilde{T}_{\text{sup}}^T \mathbb{1}^T A^T Z \tilde{T}_{\text{out}}, \quad (45)$$

and

$$\dot{\Xi}_2(\tilde{D}, \tilde{T}_{\text{out}}) = \tilde{D}^T \dot{D} = \tilde{D}^T (\frac{\mathbb{1}\mathbb{1}^T}{n} - I_n) B^T Z \tilde{T}_{\text{out}} \quad (46)$$

$$= \tilde{D}^T \frac{\mathbb{1}\mathbb{1}^T}{n} B^T Z \tilde{T}_{\text{out}} - \tilde{D}^T B^T Z \tilde{T}_{\text{out}}. \quad (47)$$

Note that $\mathbb{1}^T D(t) = D^*$ is satisfied for all $t \geq 0$. In fact, first we notice that $\mathbb{1}^T \dot{D} = 0$ at all times $t \geq 0$. Hence if $\mathbb{1}^T D(0) = D^*$ then $\mathbb{1}^T D(t) = D^*$ for all $t \geq 0$. With this we see that $\tilde{D}^T \mathbb{1} = (D - \bar{D})^T \mathbb{1} = D^* - D^* = 0$ such that (47) is reduced to

$$\dot{\Xi}_2(\tilde{D}, \tilde{T}_{\text{out}}) = -\tilde{D}^T B^T Z \tilde{T}_{\text{out}}. \quad (48)$$

Now consider the following Lyapunov function $V(\tilde{T}_{\text{out}}) = \frac{1}{2}\tilde{T}_{\text{out}}^T Z \tilde{T}_{\text{out}}$, where Z is defined in (41). Then $V(\tilde{T}_{\text{out}})$ satisfies

$$\dot{V}(\tilde{T}_{\text{out}}) = -\left\|\tilde{T}_{\text{out}}\right\|^2 - \tilde{T}_{\text{sup}}^T \mathbb{1}^T A^T Z \tilde{T}_{\text{out}} + \tilde{D}^T B^T Z \tilde{T}_{\text{out}}. \quad (49)$$

Hence, the total Lyapunov function $\Xi_1 + \Xi_2 + V$ satisfies

$$\dot{V}(\tilde{T}_{\text{out}}) + \dot{\Xi}_1(\tilde{T}_{\text{sup}}, \tilde{T}_{\text{out}}) + \dot{\Xi}_2(\tilde{D}, \tilde{T}_{\text{out}}) = -\left\|\tilde{T}_{\text{out}}\right\|^2 \leq 0. \quad (50)$$

Since $\Xi_1 + \Xi_2 + V$ is radially unbounded, (50) implies boundedness of the solutions. Using LaSalle's invariance principle this result implies that every solution to the closed loop system initialized as $\mathbb{1}^T D(0) = D^*$ converges to the largest invariant set where $\tilde{T}_{\text{out}} = 0$. Next we show that $\tilde{D}$ and $\tilde{T}_{\text{sup}}$ are zero on this invariant set. Because $\tilde{T}_{\text{out}}$ is zero, (42) reduces to

$$\mathbb{0} = -A\mathbb{1}\tilde{T}_{\text{sup}} + B\tilde{D}. \quad (51)$$

Pre-multiplying this by $\mathbb{1}^T B^{-1}$ we get

$$-\mathbb{1}^T \tilde{D} = 0 = -\mathbb{1}^T B^{-1} A \mathbb{1} \tilde{T}_{\text{sup}}, \quad (52)$$

and since

$$-\mathbb{1}^T B^{-1} A \mathbb{1} > 0, \quad (53)$$

we obtain that $\tilde{T}_{\text{sup}} = 0$. To understand why (53) holds true, observe that $A\mathbb{1}$ has all entries strictly negative, as it is immediately deduced from (63) and (64) in Appendix B. Now the inequality easily follows.

With $\tilde{T}_{\text{sup}} = 0$ and with B non-singular it follows from (51) that $\tilde{D} = \mathbb{0}$. Hence, the largest invariant set to which the solutions converge is the singleton $(\tilde{T}_{\text{out}}, \tilde{T}_{\text{sup}}, \tilde{D}) = (\mathbb{0}, 0, \mathbb{0})$. Therefore we conclude that system (42) with controllers (39) and (40) satisfies control objectives (13)-(15), and the state and the inputs of the system converge to the optimal solution. □

The proposed controller for the workload rebalances the workload currently present in the data center. The initial scheduling is assumed to be taken care of by an external entity over which we have no control. This approach is most applicable in cases where the initial scheduling is done in a non-controllable way, e.g. when the scheduling is hard-coded and incoming jobs are scheduled by means of chassis numbers.

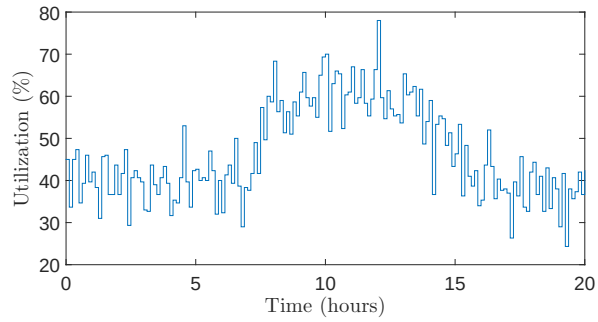

Fig. 2. Synthetic workload trace supplied to data center. The workload varies $\pm 10\%$ about two nominal values, representing nighttime and daytime operation levels. The total workload changes every 7.5 minutes during which the workload is assumed to be constant.

In these situations the only option available is to move jobs around to drive the data center to the the optimal state.

The above result shows guaranteed asymptotic tracking of constant reference signals. However in practice, the controller can handle variations in setpoints, provided that the setpoints change sufficiently slow. In the next section we will study the behavior of our controller under varying setpoints in a real data center context.

VI. Case Study

To evaluate the performance of the proposed controller, we use Matlab to simulate the closed loop system with a synthetic workload trace. For both the data center parameters and the workload trace we use the data presented in [13]. The data center parameters were obtained from measurements by Vasic et al. at the IBM Zurich Research Laboratory. This data is to our best knowledge the most extensive characterization of the heat recirculation parameters of a data center.

A. Data center parameters

The simulated data center consists of 30 homogeneous server racks, i.e. the power consumption characteristics, the safe temperature threshold and physical parameters are identical for all 30 racks. The rack model is a Dell PowerEdge 1855, with 10 dual-processor blade servers, i.e. a total of 20 CPU units per rack. The power consumption of the racks is modeled by $P_i(t) = 1728 + 145.5 D_i(t)$ [12]. The safe threshold temperature is set at 30°C. We supply a synthetic workload trace to the data center, see Fig. 2. The workload trace is constructed by varying the total workload by $\pm 10\%$ about two nominal values, 40% and 60% of the total data center capacity, representing nighttime and daytime operation levels respectively. The total workload is a piecewise constant function which changes value every 7.5 minutes. Each time the total workload changes new work is added by or released to an external entity over which we assume to have no control. After this update has taken place we observe the change in temperature from the desired temperature profile. When $(T_{\text{out}} - \bar{T}_{\text{out}})$ starts deviating from 0 the controllers will act to respond to the changing conditions.

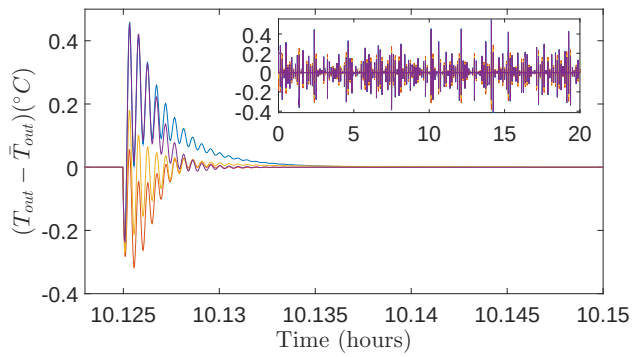

Fig. 3. Plot of the response of $(T_{\text{out}} - \bar{T}_{\text{out}})$ during the simulation for 4 selected racks. The full simulation is shown in the inset and the main plot is a magnification of the response after a change in total workload around $t = 10$ hours. Each time the total workload changes, the temperature of the racks start to deviate from the optimal value and the controllers drive the data center to the new optimal solution, $(T_{\text{out}} - \bar{T}_{\text{out}}) = \mathbb{0}$ again. The oscillatory response of the output temperature coincides with the response of the supply temperature controller. Over the whole simulation the temperature is kept in a bandwidth of ± 0.5 °C around the target temperature distribution.

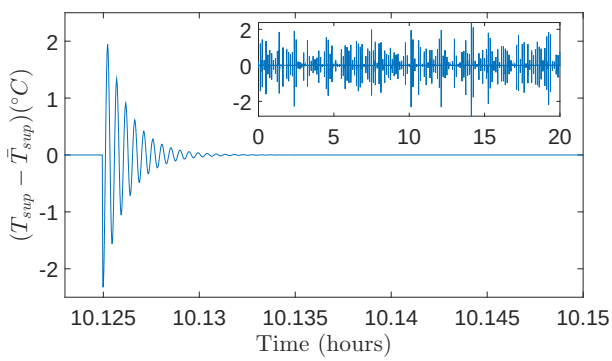

Fig. 4. Plot of the response of $(T_{\text{sup}} - \bar{T}_{\text{sup}})$ during the simulation for 4 selected racks. The full simulation is shown in the inset and the main plot is a magnification of the response after a change in total workload around $t = 10$ hours. The controller successfully drives the system to the new optimal value under varying total workload. The initial overshoot depends on the change of the total workload, i.e. the difference between the optimal supply temperatures in the two intervals. The oscillatory response results in an oscillatory fluctuation in the output temperature profile.

In Fig. 3, Fig. 4 and Fig. 5 the responses of $(T_{\text{out}} - \bar{T}_{\text{out}})$, $(T_{\text{sup}} - \bar{T}_{\text{sup}})$ and $(D - \bar{D})$ respectively for 4 selected racks are shown. To investigate the performance of the controllers we calculated the optimal values for the variables offline and used those to plot the incremental variables. The initial overshoots the Figures depend on the change in total workload between intervals. The larger the change, the larger this initial overshoot will be. We observe different behavior for the two controllers. The controller for the supply temperature results in very oscillatory behavior for the supply temperature which in turn results in a fluctuating output temperature profile. The controller for the workload division however shows a much smoother response and more gradually steers the workload distribution to the optimal distribution. Every time the workload changes the controllers drive the system back to the optimal value in approximately 0.01 hour = 36 seconds.

In Fig. 6 the response of $(T_{\text{out}} - \bar{T}_{\text{out}})$ is shown for a larger

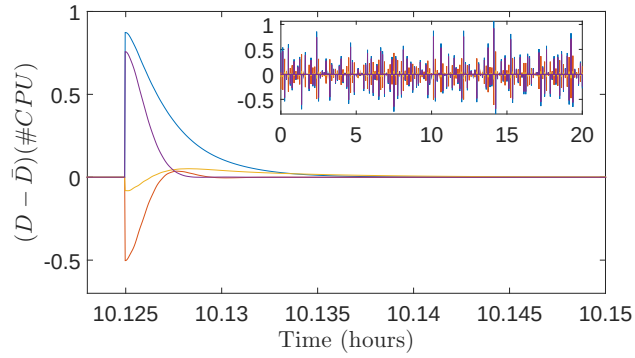

Fig. 5. Plot of the response of $(D - \bar{D})$ during the simulation for 4 selected racks. The full simulation is shown in the inset and the main plot is a magnification of the response after a change in total workload around $t = 10$ hours. As above the controller drives the system to the optimal value each time the total workload changes. When the total workload changes, an external entity adds or subtracts work from the racks in a non-optimal way which causes an initial overshoot. The controller redistributes the work again to the new optimal workload distribution.

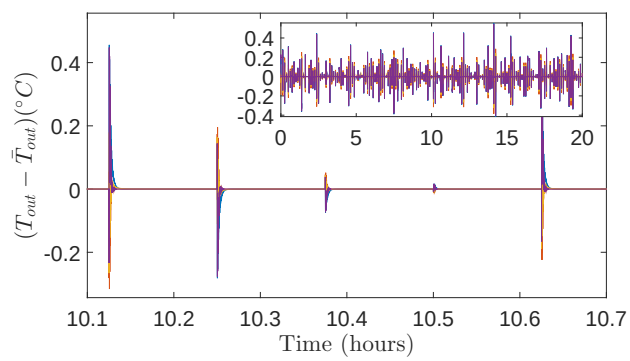

Fig. 6. Plot of the response of $(T_{\text{out}} - \bar{T}_{\text{out}})$ during the simulation for 4 selected racks. The full simulation is shown in the inset and the main plot shows the temperature response over a larger time interval which covers multiple changes in total workload. The fast response of the controllers is clearly visible here and we see that, after a very short transient, the controllers steer the temperature of the servers back to the optimal value.

time interval. In this time interval the total workload changes multiple times and it is seen how, after a very short transient, the controllers steer the temperature of the servers back to the optimal value. This shows that our controllers can cope with variations in total workload.

Although this is a very quick response it is not likely that this convergence time will be attained in practice. In the simulation the cooled air of the CRAC instantly reaches the racks, whereas in a real data center it will take some time for the air to travel from the CRAC to the racks. On the contrary the workload division happens on a much shorter timescale, therefore we expect that in practice the output temperature will first increase, as new work is assigned to the rack, and after a certain delay the cooling will start to kick in to drive the temperature profile back to the setpoint.

The supplied workload simulated a day and night cycle to study the response of the controller under large varying loads. From the results we see no difficulty for the controller to handle these different conditions. We conclude that the

controller is able to keep the temperature of the racks around the target setpoint under all load conditions.

VII. Conclusions and Future Work

Many papers on thermal-aware job scheduling have studied the topic from a practical perspective, however a theoretical analysis has less often been done. In this work we describe data centers and corresponding thermodynamics in a control theoretical fashion combining optimization theory with controller design.

We have studied the minimization of energy consumption in a data center where recirculation of airflow is present, i.e. inefficiencies in cooling of the racks, through thermal-aware job scheduling and cooling control. We have set up an optimization problem and characterized the optimal workload distribution and cooling temperature to achieve minimum energy consumption while ensuring job processing and thermal threshold satisfaction. In addition we have presented controllers that track a reference signal and are able to drive the control and state variables to the optimal values. Furthermore simulations show that the controllers can work with varying workload conditions as the convergence time of the controllers is significantly faster than the frequency of the workload variation.

We have shown that it is possible to uniquely determine the optimal cooling supply temperature and workload distribution as a function of the total workload and desired temperature distribution of the racks in the data center. Furthermore we have shown that the optimal temperature distribution can be analytically calculated and that this distribution is independent of the workload distribution if none of the racks reaches its computational capacity.

With the assumption that none of the racks is at its computational capacity we have designed controllers that control the supply temperature and workload distribution to drive the data center to the optimal state.

There are several directions in which we want to extend our research. First we want to extend the framework to include situations where the optimal temperature distribution changes due to racks reaching their computational capacity. This will allow us to include server consolidation where the number of active racks is decreased to further reduce energy consumption. In these situations it is inevitable that the computational capacity of some of the racks is reached and that varying optimal temperature distributions will have to be addressed.

Our control approach requires knowledge of the thermal characteristics of the data center. Studying the robustness and stability of our approach under small variations of the heat recirculation matrix is therefore of importance. Lastly it would be interesting to study the possibility of allowing multiple CRAC units with different set points, or including other variables in the optimization problem, such as Service Level Agreements and response times of the jobs.

Appendix A
Proof of Property 1

From Lemma 3 we have that

$$C_3 = -W^{-1}MA(I_n - \mathbb{1}C_1^T),$$

where

$$C_1^T = \frac{\mathbb{1}^T W^{-1} MA}{\mathbb{1}^T W^{-1} MA \mathbb{1}}.$$

Defining a temporary variable $\alpha = W^{-1}MA$ we can write C_3 as

$$C_3 = -\alpha + \frac{1}{\mathbb{1}^T \alpha \mathbb{1}} \alpha \mathbb{1}\mathbb{1}^T \alpha.$$

The ij-th component of C_3 is then given by

$$C_3^{ij} = -\alpha_{ij} + \frac{\sum_{l=1}^n \alpha_{il} \sum_{k=1}^n \alpha_{kj}}{\sum_{l=1}^n \sum_{k=1}^n \alpha_{lk}}. \tag{54}$$

From the definition of α we find that the ij-th component of α is given by

$$\alpha_{ij} = c_p \rho \frac{1}{w_i} (\gamma_{ji} - \delta_{ji}) f_j, \tag{55}$$

where δ_{ji} is the Kronecker delta, which is 1 if $i = j$ and 0 otherwise. To simplify the mathematics a little from now on, we assume that the data center consists of homogeneous racks, see (18). Combining (55) with (54) we have

$$\begin{aligned} C_3^{ij} = &- c_p \rho \frac{1}{w} \Bigg((\gamma_{ji} - \delta_{ji}) f_j \\ &+ \frac{(f_i - \sum_{l=1}^n \gamma_{li} f_l)(f_j - \sum_{k=1}^n \gamma_{jk} f_j)}{\sum_{l=1}^n (f_l - \sum_{k=1}^n \gamma_{kl} f_k)} \Bigg). \end{aligned} \tag{56}$$

Although the big fraction in (56) looks a bit daunting it is actually easy to conceptually understand it. The airflow at the inlet of the rack consists of two parts, air coming from the CRAC unit and air recirculating from other racks to the rack in question. At the outlet of the rack the airflow is composed of the air going back to the CRAC unit and the air recirculating from the rack in question to all the other racks. Looking closer at the numerator of (56) we see that the first half is the air flowing from the CRAC unit to rack i, and the second half is the air flowing from rack j to the CRAC unit. The denominator is the sum of the airflow each rack receives from the CRAC unit which is equal to the supplied airflow, f_{sup}. In this way we can simplify (56) to

$$C_3^{ij} = -c_p \rho \frac{1}{w} \left((\gamma_{ji} - \delta_{ji}) f_j + \frac{f_{(\text{CRAC to i})} f_{(\text{j to CRAC})}}{f_{\text{sup}}} \right). \tag{57}$$

Now in the case that $i \neq j$, (57) is reduced to

$$C_3^{ij} = \underbrace{-c_p \rho \frac{1}{w}}_{<0} \left(\underbrace{\gamma_{ji} f_j}_{>0} + \underbrace{\frac{f_{(\text{CRAC to i})} f_{(\text{j to CRAC})}}{f_{\text{sup}}}}_{>0} \right) < 0. \tag{58}$$

Here we see that the off-diagonal terms of C_3 are strictly negative.

As for the diagonal terms, $i = j$, we have

$$C_3^{ii} = c_p \rho \frac{1}{w} \left((1 - \gamma_{ii}) f_i - \frac{f_{(\text{CRAC to i})} f_{(\text{i to CRAC})}}{f_{\text{sup}}} \right). \tag{59}$$

Since

$$(1 - \gamma_{ii}) f_i = \underbrace{f_i - \sum_{l=1}^n \gamma_{li} f_l}_{f_{(\text{CRAC to i})}} + \sum_{l=1, l \neq i}^n \gamma_{li} f_l, \tag{60}$$

we have that

$$C_3^{ii} = \underbrace{c_p \rho \frac{1}{w}}_{>0} \Bigg(\underbrace{\sum_{l=1, l\neq i}^{n} \gamma_{li} f_l}_{>0} + \underbrace{f_{(\text{CRAC to i})}}_{>0} \underbrace{\left(1 - \frac{f_{(\text{i to CRAC})}}{f_{\text{sup}}}\right)}_{>0} \Bigg) > 0. \tag{61}$$

In (61) we see that the diagonal terms of C_3 are strictly positive. This concludes the proof. □

Appendix B
Proof of Hurwitz property of matrix A

Matrix A as defined in Subsection II-C is given by

$$A = \rho c_p M^{-1} (\Gamma^T - I_n) F. \tag{62}$$

Writing the matrix out in full gives

$$A = \rho \begin{pmatrix} \frac{\gamma_{11}-1}{m_1} f_1 & \frac{\gamma_{21}}{m_1} f_2 & \cdots & \frac{\gamma_{n1}}{m_1} f_n \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\gamma_{1n}}{m_n} f_1 & \frac{\gamma_{2n}}{m_n} f_2 & \cdots & \frac{\gamma_{nn}-1}{m_n} f_n \end{pmatrix}. \tag{63}$$

If we can show that matrix A is strictly diagonal dominant and that the diagonal elements are negative then by the Gerschgorin circle theorem we have shown that matrix A is Hurwitz.

First we will prove strict diagonal dominance of matrix A. As stated in Appendix A, the airflow in a rack consists of two parts, the recirculated air from the other racks and the supplied air by the CRAC, namely

$$f_i = \gamma_{ii} f_i + \sum_{j=1, j\neq i}^{n} \gamma_{ji} f_j + f_{\text{sup}}^i.$$

Hence,

$$\begin{aligned} (\gamma_{ii} - 1) f_i &= - \sum_{j=1, j\neq i}^{n} \gamma_{ji} f_j - f_{\text{sup}}^i \\ &< - \sum_{j=1, j\neq i}^{n} \gamma_{ji} f_j, \end{aligned} \tag{64}$$

from which

$$|(\gamma_{ii} - 1) f_i| > \left| - \sum_{j=1, j\neq i}^{n} \gamma_{ji} f_j \right| = \sum_{j=1, j\neq i}^{n} \gamma_{ji} f_j, \tag{65}$$

because all γ_{ij} are strictly between 0 and 1. Comparing (65) with (63) and ignoring the mass, as the same mass appears in every row i, we see that matrix A is strictly diagonal dominant.

Furthermore as γ_{ii} is strictly between 0 and 1, we have that all the diagonal elements of A are strictly negative. By Gerschgorin circle theorem, all the eigenvalues of matrix A are strictly negative and therefore the matrix is Hurwitz. □

Acknowledgment

This research was carried out as part of the perspective program Robust Design of Cyber-Physical Systems, Cooperative Networked Systems and is supported by Technology Foundation STW and industrial partners Target Holding and Better.be. The authors would also like to thank IBM Zurich Research Lab for supplying measurement data of a real-life data center.

References

[1] T. Van Damme, C. De Persis, and P. Tesi, "Optimized thermal-aware job scheduling and control of data centers," in *Proceedings of the IFAC World Congress*, 2017.

[2] D. Blatch, "Is the industry getting better at using power?" *Datacenter Dynamics Focus*, vol. 3, no. 33, pp. 16–17, Jan/Feb 2014.

[3] Enerdata, "Global domestic electricity consumption," August 2016.

[4] A. Hameed, A. Khoshkbarforoushha, R. Ranjan, P. P. Jayaraman, J. Kolodziej, P. Balaji, S. Zeadally, Q. M. Malluhi, N. Tziritas, A. Vishnu, S. U. Khan, and A. Zomaya, "A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems," *Computing*, pp. 1–24, jun 2014.

[5] J. Moore, J. Chase, R. Parthasarathy, and S. Ratnesh, "Making scheduling 'cool' temperature-aware workload placement in data centers," in *USENIX Annual Technical Conference*, 2005, pp. 61–74.

[6] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, "Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: a cyber-physical approach," *IEEE trans. on Parallel and distributed systems*, vol. 19, pp. 1458–1472, nov 2008.

[7] H. Sun, P. Stolf, J.-M. Pierson, and G. Da Costa, "Energy-efficient and thermal-aware resource management for heterogeneous datacenters," *Sustainable Computing: Informatics and Systems*, vol. 4, no. 4, pp. 292–306, 2014.

[8] X. Jiang, M. I. Alghamdi, M. M. Al Assaf, X. Ruan, J. Zhang, M. Qiu, and X. Qin, "Thermal modeling and analysis of cloud data storage systems," *Journal of Communications*, vol. 9, pp. 299–311, Apr 2014.

[9] A. Pahlavan, M. Momtazpour, and M. Goudarzi, "Power reduction in hpc data centers: a joint server placement and chassis consolidation approach," *The Journal of Supercomputing*, vol. 70, no. 2, pp. 845–879, 2014.

[10] E. Pakbaznia, M. Ghasemazar, and M. Pedram, "Temperature-aware dynamic resource provisioning in a power-optimized datacenter," in *Proceedings of the Conference on Design, Automation and Test in Europe*. European Design and Automation Association, 2010, pp. 124–129.

[11] S. Li, H. Le, N. Pham, J. Heo, and T. Abdelzaher, "Joint optimization of computing and cooling energy: Analytic model and a machine room case study," in *32nd Int. Conf. on Distributed Computing Systems*. IEEE, Jun 2012, pp. 396–405.

[12] Q. Tang, S. K. S. Gupta, D. Stanzione, and P. Cayton, "Thermal-aware task scheduling to minimize energy usage of blade server based datacenters," in *2nd IEEE Int. Symp. on Dependable, Autonomic and Secure Computing*. IEEE, Sep 2006, pp. 195–202.

[13] N. Vasic, T. Scherer, and W. Schott, "Thermal-aware workload scheduling for energy efficient data centers," in *Proceedings of the 7th international conference on Autonomic computing*, 2010, pp. 169–174.

[14] X. Yin and B. Sinopoli, "Adaptive robust optimization for coordinated capacity and load control in data centers," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*. IEEE, 2014, pp. 5674–5679.

[15] J. Doyle, F. Knorn, D. O'Mahony, and R. Shorten, "Distributed thermal aware load balancing for cooling pf modular data centres," *IET Control Theory and Applications*, vol. 7, pp. 612–622, Nov 2013.

[16] C. Albea, A. Seuret, and L. Zaccarian, "Hybrid control of a three-agent network cluster," in *53th IEEE Conference on Decision and Control*, dec 2014, pp. 5302–5307.

[17] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. Wang, "A cyber-physical systems approach to data center modeling and control for energy efficiency," *Proceedings of the IEEE*, vol. 100, pp. 254–268, jan 2012.

[18] M. Bürger and C. De Persis, "Dynamic coupling design for nonlinear output agreement and time-varying flow control," *Automatica*, vol. 51, pp. 210–222, 2015.

[19] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 2016.

[20] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ACM SIGARCH Computer Architecture News*, vol. 35. ACM, 2007, pp. 13–23.

[21] B. E. Lauri Minas, "The problem of power consumption in servers," Intel, Santa Clara, CA, USA, Tech. Rep., 2009.

[22] V. Gupta, R. Nathuji, and K. Schwan, "An analysis of power reduction in datacenters using heterogeneous chip multiprocessors," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 87–91, 2011.

[23] Q. Tang, T. Mukherjee, S. K. S. Gupta, and P. Cayton, "Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters," in *Fourth International Conference on Intelligent Sensing and Information Processing, 2006. ICISIP 2006*. IEEE, December 2006, pp. 203–208.

[24] T. Heath, A. P. Centeno, P. George, L. Ramos, Y. Jaluria, and R. Bianchini, "Mercury and freon: temperature emulation and management for server systems," in *12th international conference on Architectural support for programming languages and operating systems*. ASPLOS XII, 2006, pp. 106–116.

[25] P. Ranganathan, P. Leech, I. David, and C. J. S., "Ensemble-level power management for dense blade servers," in *33rd annual international symposium on Computer Architecture*. ISCA, 2006, pp. 66–77.

[26] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.