Optimized Thermal-Aware Job Scheduling and Control of Data Centers

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

Published in:
Proceedings of the IFAC 2017 conference

DOI:
10.1016/j.ifacol.2017.08.1393

Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 50-1 (2017) 8244–8249

Optimized Thermal-Aware Job Scheduling and Control of Data Centers

Tobias Van Damme * **Claudio De Persis** * **Pietro Tesi** *

* *University of Groningen, 9747 AG Groningen, The Netherlands (e-mail: {t.van.damme, c.de.persis, p.tesi}@rug.nl).*

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 drive the data center 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.

Keywords: Optimization and control of large-scale network systems; Networked systems; Lyapunov methods; Control of constrained systems; Cyber-Physical Systems

1. INTRODUCTION

Data centers are big energy consumers, in 2013 data centers consumed 350 billion kWh of energy, 1.73% of the global electricity consumption (Blatch, 2014; Enerdata, 2016). 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 (Hameed et al., 2014).

Although much progress has been made, there are still several challenges 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 have studied strategies which uses the knowledge of the thermal properties of the data center to make more intelligent choices how to schedule incoming jobs (Moore et al., 2005; Tang et al., 2008). With heuristic methods they showed improvements of up to 30% less energy consumption with respect to non thermal-aware job schedulers. On the other hand, studies have also been done in a more theoretical direction. Cast as a control problem (Vasic et al., 2010) has proposed a control algorithm that tries to maintain the temperature of the equipment around a target value. In (Yin and Sinopoli, 2014) a two-step algorithm is proposed 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.

While all this work has strong points on its own, to the authors best knowledge a thorough analysis and characterization of an energy minimal solution combined with a straightforward control strategy which handles both cooling and job scheduling simultaneously has not been done before. The objective of this work is to supply an easily extendable framework that allows for a characterization of an energy-minimal operating point and then supply straightforward methods for operating the data center such that this operating point is achieved for all load conditions. In addition it should be extendable to include more complex concepts, like switching on and off servers or including quality-of-service constraints.

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 combines energy minimization with the thermodynamics. In addition to only including temperature constraints (Li et al., 2012) 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.

Secondly 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 (Vasic et al., 2010; Parolini et al., 2012). However in (Vasic et al., 2010) the two control goals were handled separately; In (Parolini et al., 2012) 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 which is guaranteed to converge to the energy minimal solution. Our method is inspired by results from (Bürger and De Persis, 2015) where regulation to optimal steady solutions in the presence of disturbances was considered. Therefore our strategy also allows for vary-

⋆ The authors declare no competing financial interest

10.1016/j.ifacol.2017.08.1393

ing and unknown workload changes while guaranteeing convergence to the energy-minimal operating point.

The remainder of this paper is organized as follows. In Section 2 the basic thermodynamics are formulated. Then an optimization problem is formulated in Section 3 and the equivalence to a reduced form is proven. Following up, the optimal solution is analytically analyzed and characterized for different load conditions in Section 4. Using this analytical solution a controller is proposed in Section 5 that can handle unknown load conditions. Finally in Section 6 a case study is considered to show the performance of the controllers.

Due to space constraints, all the proofs can be found in (Van Damme et al., 2016).

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_3^{ij}. We write the diagonal matrix constructed from the elements of vector x as $\text{diag}\{x_1,\, x_2,\, \cdots,\, x_n\}$. 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 $\mathbf{0} \in \mathbb{R}^n$. 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. Finally a data center consists of n racks.

2. 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.

2.1 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. This variable is collected in the vector

$$D(t) := \left(D_1(t)\ D_2(t)\ \cdots\ D_n(t)\right)^T.$$

2.2 Power consumption of racks

The most common way to model the power consumption of a single rack is using a linear model (Heath et al., 2006). In this way the power consumption, $P_i(t)$, of a rack is modeled 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

$$P(t) := \left(P_1(t)\ P_2(t)\ \cdots\ P_n(t)\right)^T,$$
$$V := \left(v_1\ v_2\ \cdots\ v_n\right)^T,$$

and

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

so that

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

2.3 Thermodynamical model

Following similar arguments as in (Vasic et al., 2010) a thermodynamical model for each individual rack is constructed. 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

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

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

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

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

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

Here

$$T_{\text{out}}(t) := \left(T^1_{\text{out}}(t)\ T^2_{\text{out}}(t)\ \cdots\ T^n_{\text{out}}(t)\right)^T,$$

and

$$\begin{aligned} A &:= \rho c_p M^{-1}(\Gamma^T - I_n)F, \\ F &:= \text{diag}\{f_1,\ f_2,\ \cdots,\ f_n\}, \\ M &:= \text{diag}\{c_pm_1,\ c_pm_2,\ \cdots,\ c_pm_n\}, \\ \Gamma &:= [\gamma_{ij}]_{n\times n}. \end{aligned}$$

2.4 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

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

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

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

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

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

To determine the amount of work the CRAC has to do to remove a certain amount of heat, Moore et al. (Moore et al., 2005) 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. In a general sense the COP can be any monotonically increasing function. The power consumption of the CRAC units can then be given by

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

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

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

3. 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

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

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

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

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 3.1. ■

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

3.1 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

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

A reasonable way (Li et al., 2012; Yin and Sinopoli, 2014) to formulate the optimization problem is

$$\begin{aligned} \min_{T_{\text{out}},T_{\text{sup}},D}\quad & \frac{Q_{\text{rem}}}{\text{COP}(T_{\text{sup}})} + \mathbb{1}^TP(D) && (17a)\\ s.t.\quad & D^* = \mathbb{1}^TD && (17b)\\ & \mathbf{0} \preccurlyeq D \preccurlyeq D_{\max} && (17c)\\ & \mathbf{0} = A(T_{\text{out}} - \mathbb{1}T_{\text{sup}}) + M^{-1}P(D) && (17d)\\ & T_{\text{out}} \preccurlyeq T_{\text{safe}}. && (17e) \end{aligned}$$

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, $T_{\text{safe}} \in \mathbb{R}^n$.

3.2 Reduced optimization problem

Due to the non-linear nature of how the COP affects the power consumption it is not trivial to analyze this

problem. However under some mild assumptions it is possible to reduce the optimization defined in (17) to a simpler problem.

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

$$\max_{T_{\text{out}}} \quad C_1^T T_{\text{out}} \tag{18a}$$
$$s.t. \quad \mathbf{0} \preccurlyeq C_3 T_{\text{out}} + C_4(D^*) \preccurlyeq D_{\max} \tag{18b}$$
$$T_{\text{out}} \preccurlyeq T_{\text{safe}}, \tag{18c}$$

for suitable C_1, C_3 and C_4. ■

For understanding this theorem we introduce some notation and extra theory.

Lemma 1. Equations (11) and (17d) imply that the following relation holds

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

which reduces the cost function to

$$\mathcal{C}(T_{\text{out}}, T_{\text{sup}}, D) = \left(1 + \frac{1}{\text{COP}(T_{\text{sup}})}\right) \mathbb{1}^T P(D). \tag{19}$$

■

Remark 1. In many real-life data centers most of the equipment is identical, i.e. such that $v_i = v$ and $w_i = w$ for all i in (2). In this case the data center is said to be homogeneous and its power consumption is given by $P(D) = v\mathbb{1} + wD$. The total computational power consumption is then given by

$$\mathbb{1}^T P(D) = nv + w\mathbb{1}^T D = nv + wD^*. \tag{20}$$

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 (19) considerably. ■

Lemma 2. If (17b) and (17d) are satisfied, then there is a unique supply temperature which follows from the desired, chosen temperature distribution, namely

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

■

Lemma 3. If (17b) and (17d) are satisfied, then there is a unique workload distribution which follows from the desired, chosen temperature distribution, i.e.

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

■

Remark 2. The dimensions of the constants from above lemma's 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

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

■

Lemma 2 and Lemma 3 show that at the steady state the supply temperature and workload distribution are uniquely defined by the total workload, D^*, and the temperature distribution, T_{out}.

4. 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.

4.1 KKT optimality conditions

Because the optimization problem (18) 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 (18) 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:

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

4.2 Optimal solution for output temperature

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

The characterization of 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.

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

The solution to (24) and the optimal solution for the optimization problem (18) is then given by

$$\bar{\mu}_+ = \bar{\mu}_- = \mathbf{0}, \quad \bar{\mu} = C_1 \succ \mathbf{0}, \quad \bar{T}_{\text{out}} = T_{\text{safe}}. \tag{25}$$

■

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

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

the solution of (24) is as follows:

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

$$\bar{\mu}^i_- = 0, \quad \frac{C^i_1 + \sum_{j=1,j\neq i}^{k} \bar{\mu}^j_+ \left|C^{ji}_3\right|}{C^{ii}_3} \geq \bar{\mu}^i_+ \geq 0, \tag{26}$$

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

$$\begin{aligned}\bar{T}^i_{\text{out}} &= \frac{D^i_{\max} - C^i_4(D^*)}{C^{ii}_3} + \sum_{j=k+1}^{n} \frac{\left|C^{ij}_3\right|}{C^{ii}_3} T^j_{\text{safe}} \\ &\quad + \sum_{j=1,j\neq i}^{k} \frac{\left|C^{ij}_3\right|}{C^{ii}_3} \bar{T}^j_{\text{out}} \\ &\leq T^i_{\text{safe}}.\end{aligned} \tag{28}$$

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

$$\bar{\mu}^i_- = \bar{\mu}^i_+ = 0, \tag{29}$$

$$\bar{\mu}^i = C^i_1 + \sum_{j=1}^{k} \bar{\mu}^j_+ \left|C^{ji}_3\right| > 0, \tag{30}$$

$$\bar{T}^i_{\text{out}} = T^i_{\text{safe}}. \tag{31}$$

■

5. TEMPERATURE BASED JOB SCHEDULING CONTROL

As established in Section 4 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 $\mathbf{0} \prec D \prec D_{\max}$. 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.

5.1 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

$$\dot{T}_{\text{sup}} = \mathbb{1}^T A^T Z (T_{\text{out}} - T_{\text{safe}}), \tag{32}$$

$$\dot{D} = (\frac{\mathbb{1}\mathbb{1}^T}{n} - I_n) B^T Z (T_{\text{out}} - T_{\text{safe}}), \tag{33}$$

where A is Hurwitz, Z is the symmetric positive definite matrix such that

$$A^T Z + Z A = -2 I_n, \tag{34}$$

and B is

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

where W is defined Subsection 2.2, and A and M are defined in Subsection 2.3. The controllers (32) and (33) 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 theorem below we study the convergence behavior of the controllers in a time frame where the total workload, D^*, is assumed to be constant.

Theorem 4. Assume D^* is constant and $\mathbb{1}^T D(0) = D^*$. Then the solution of system (8) with controllers (32) and (33) is bounded and converges to the optimal solution of the optimal problem defined in (17) and therefore satisfies control objectives (13)-(15). ■

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 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.

6. 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 (Vasic et al., 2010). 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.

6.1 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. The power consumption of the racks is modeled by $P_i(t) = 1728 + 145.5 D_i(t)$ (Tang et al., 2006). The safe threshold temperature is set at 30°C. We supply a synthetic workload trace to the data center, see Fig. 1. The workload trace is constructed by varying the total workload by ±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.

In Fig. 2 the response of $(T_{\text{out}} - \bar{T}_{\text{out}})$ for 4 selected racks is shown. To investigate the performance of the controllers we calculated the optimal values for the variables offline

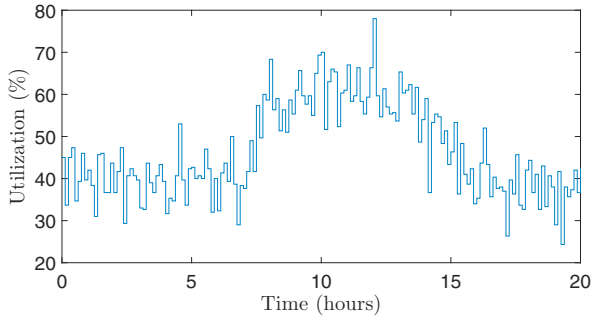

Fig. 1. Synthetic workload trace supplied to data center. The workload varies ±10% about two nominal values, representing nighttime and daytime operation levels. The total workload changes every 7.5 minutes during which the workload is assumed to be constant.

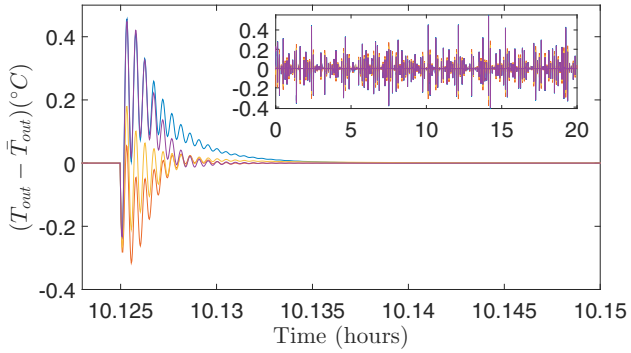

Fig. 2. Plot of the response of $(T_{\text{out}} - \bar{T}_{\text{out}})$ during the simulation for 4 selected racks. The full simulation is shown in the inset and the main plot is a magnification of the response after a change in total workload around $t = 10$ hours.

and used those to plot deviation from the optimal values. The initial change in Fig. 2 is dependent 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. Every time the workload changes the controllers drive the system back to the optimal value in approximately 0.01 hour = 36 seconds.

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.

7. 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 a controller that works under varying workload conditions and is able to drive the control and state variables to the optimal values.

An interesting direction 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 reduce energy consumption. In these situations it is inevitable that the computational capacity of the racks is reached and that varying optimal temperature distributions will have to be addressed.

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

Blatch, D. (2014). Is the industry getting better at using power? *Datacenter Dynamics Focus*, 3(33), 16–17.

Bürger, M. and De Persis, C. (2015). Dynamic coupling design for nonlinear output agreement and time-varying flow control. *Automatica*, 51, 210–222.

Enerdata (2016). Global domestic electricity consumption.

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

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

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

Moore, J., Chase, J., Parthasarathy, R., and Ratnesh, S. (2005). Making scheduling 'cool' temperature-aware workload placement in data centers. In *USENIX Annual Technical Conference*, 61–74.

Parolini, L., Sinopoli, B., Krogh, B.H., and Wang, Z. (2012). A cyber-physical systems approach to data center modeling and control for energy efficiency. *Proceedings of the IEEE*, 100, 254–268.

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

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

Van Damme, T., De Persis, C., and Tesi, P. (2016). Optimized thermal-aware job scheduling and control of data centers. *arXiv:1611.00522*.

Vasic, N., Scherer, T., and Schott, W. (2010). Thermal-aware workload scheduling for energy efficient data centers. In *Proceedings of the 7th international conference on Autonomic computing*, 169–174.

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