Preprints of the 18th IFAC World Congress
Milano (Italy) August 28 - September 2, 2011

Model Predictive Control of Data Centers in the Smart Grid Scenario

Luca Parolini, Bruno Sinopoli, Bruce H. Krogh

Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA 15213-3890, USA
(e-mail: {lparolin—brunos—krogh}@ece.cmu.edu)

Abstract: This paper discusses the challenges and the possibilities offered by controlling a data center as a node of the smart-grid. Communication between the grid and a data center takes the form of a time-varying and power-consumption-dependent electricity price. A cyber-physical approach is considered. The computational and the physical characteristics of a data center, as well as the interactions between the two aspects, are explicitly represented. Simulation results show that the proposed control approach can lead to larger income for data center operators than other approaches that do not consider the interactions between the cyber and the physical subsystems.

Keywords: Analysis and control in deregulated power systems; Smart grids; Data centers; Model Predictive control.

1. INTRODUCTION

Power consumption of data centers has increased significantly in the past few years [U.S. Environmental Protection Agency (2007)]. As a consequence, efficiently powering and cooling data centers has become a challenging problem. Modern large-scale data centers are tailored for peak power consumption of dozens of MW and can have operational costs as high as $5.6M [Fan et al. (2007); Hamilton (2008)]. From the grid perspective, the distinctive features of data centers relative to other electricity costumers are their high value of power consumption per squared meter, e.g., 1.6 KW/m^2 for a high-heat-density data centers [Sharma et al. (2005)], and the time scale at which power consumption can be controlled, e.g., on the minute time-scale.

In a deregulated electricity market, the electricity price varies over time and geographic locations. In the U.S., two markets can be used to purchase electricity: the *day-ahead market* and the *real-time market*. A possible approach to reduce the variability of the electricity cost and also to tackle critical situations such as network failure is to require some costumers to cap their power consumption upon request from the grid. Such a technique, already applied by some independent system operators (ISOs), is called *demand response program* (DRP). [1]

We consider the case where a data center buys electricity on the day-ahead market and also participates in a DRP. We focus on the case where the cost of electricity depends on the amount of power the data center consumes. As long as the power consumption is lower than a time-varying threshold, the electricity price is kept at a reduced value. When the data center power consumption exceeds the given threshold, the additional power is provided at a higher cost. To maximize profit, a data center controller has to predict the future power consumption of the data center on the minute time-scale and decide whether it is more profitable to drop or delay the execution of user requests, or to buy energy at higher prices and process all of the workload.

A data center can leverage over two *service level agreements* (SLAs): SLA$_U$ with the data center users and SLA$_G$ with the grid. SLA$_U$ determines the income for the data center when it executes the required workload with a certain quality of service (QoS), e.g., within a certain delay. SLA$_G$ determines the cost of electricity over time and for different levels of power consumption.

2. PROBLEM STATEMENT

We consider a hierarchical control strategy, where a central coordinator provides a collection of bounds and optimal set points to lower-level controllers that then operate independently from each other [Parolini et al. (2010b)]. The controller at the highest level of the hierarchy is called *data center controller* and it is the subject of this paper.

Throughout the paper, the variables k and ν are elements of $\mathbb{Z}$ with $\nu \geq k$, and the variables t and τ are elements of $\mathbb{R}$ with $\tau \geq t$. With $\hat{x}(\nu|k)$ we denote the expected value of the variable $x(\nu)$ based on the information available up to the beginning of the k^{th} interval. With $\hat{x}(\tau|t)$ we denote the expected value of the variable $x(\tau)$ based on the information available up to time t.

2.1 Computational model

Multiple servers are modeled as a single computational entity called a *server node*. Every virtual machine (VM) is modeled as a queuing system, where the queued entities are called *jobs*. Jobs are divided into classes and jobs in the same class have identical resource requirements. A predictor provides estimations of future job arrival rates. The migration of a VM is approximated with the migration of the jobs contained in the VMs and it is assumed that the

⋆ This material is based upon work partially supported by the National Science Foundation under Grant No. 0925964. The authors would like to thank Zhikui Wang, Cullen E. Bash, Yuan Chen, and Daniel Gmach of Hp Labs for the helpful discussions on modeling and control of data centers.

[1] http://pjm.com/markets-and-operations/demand-response.aspx

controllers in the lower level of the hierarchy choose which VM should be migrated. Since the data center controller operates on a dozen-of minutes time-scale, the time to migrate the VMs is neglected.

Let N be the number of server nodes in the data center and J be the number of job classes. With $l_i^j(k)$ we denote the number of jobs in class j at the i^{th} server node at the beginning of the k^{th} interval. The number of jobs in class j left on the i^{th} server after the VM migration is denoted by $\tilde{l}_i^j(k)$. The variables $\{\tilde{l}_i^j(k)\}$ are constrained by

$$0 \leq \tilde{l}_i^j(k), \qquad \sum_{i=1}^{N} \tilde{l}_i^j(k) \leq \sum_{i=1}^{N} l_i^j(k). \tag{1}$$

The second inequality in (1) implies that some of the jobs being executed on a server node can be dropped (intentionally) during migration.

The number of jobs in class j generated by user requests during the k^{th} interval is denoted by $a^j(k)$ (arrival). The relative amount of jobs in class j sent to the i^{th} node during the k^{th} interval is a controllable variable denoted by $s_i^j(k)$. The actual number of jobs sent to the i^{th} node is $a_i^j(k)$. For all $i = 1, \ldots, N$ and $j = 1, \ldots, J$ we can write

$$\hat{a}_i^j(\nu|k) = \hat{s}_i^j(\nu|k)\hat{a}^j(\nu|k),$$

where $0 \leq \hat{s}_i^j(\nu|k)$, and $\sum_{i=1}^{N} \hat{s}_i^j(\nu|k) \leq 1$.

The hardware resources of a server node are divided into H classes and the amount of resources of class h available at the i^{th} node is denoted by $\overline{\Pi}_{i,h}$. A linear relationship is considered between the average number of jobs processed by a node and the average amount of hardware resources used to process the jobs. Let b_h^j be the amount of hardware resources of class h that a single job in class j requires and let $d_i^j(k)$ (departure) denote the number of jobs in class j that leaves the i^{th} node during the k^{th} interval. The average amount of hardware resources in class h used by jobs in class j is given by $d_i^j(k)b_h^j$.

The amount of hardware resources assigned to a VM is a controllable parameter. Let $\rho_i^j(k)$ denote the relative amount of hardware resources allocated by the i^{th} server node for jobs in class j during the k^{th} interval. Variables $\{\rho_i^j(k)\}$ take values in $[0, 1]$. The average amount of hardware resources in class h allocated by the i^{th} server node for jobs in class j during the k^{th} interval is $\tilde{b}_{i,h}^j \rho_i^j(k)$ and the maximum amount of jobs that a node can process is denoted by $\overline{\mu}_i^j \rho_i^j(k)$. The coefficients $\{\tilde{b}_{i,h}^j\}$ are a scaled version of the coefficients $\{b_h^j\}$. The variables $\{\rho_i^j(k)\}$ are constrained by

$$0 \leq \rho_i^j(k) \leq 1, \qquad \sum_{j=1}^{J} \rho_i^j(k)\tilde{b}_{i,h}^j \leq \overline{\Pi}_{i,h}. \tag{2}$$

The average amount of resources required to process all of the jobs during the k^{th} interval is $(\tilde{l}_i^j(k) + a_i^j(k))b_h^j$. When a server node allocates more than the average required resources to jobs in class j, then the number of jobs departing the node equals $\tilde{l}_i^j(k) + a_i^j(k)$. When a server node allocates less than the average required resources to jobs in class j, then the number of jobs departing the node equals $\overline{\mu}_i^j \rho_i^j(k)$. Thus, for all $i = 1, \ldots, N$, and $j = 1, \ldots, J$ we can write

$$\hat{d}_i^j(\nu|k) = \min\left\{\hat{\tilde{l}}_i^j(\nu|k) + \hat{a}_i^j(\nu|k),\ \overline{\mu}_i^j \rho_i^j(\nu|k)\right\}. \tag{3}$$

Let us define the vectors $\boldsymbol{l}(k) = \left[l_1^1(k)\ \ldots\ l_N^J(k)\right]^T$, $\tilde{\boldsymbol{l}}(k) = \left[\tilde{l}_1^1(k)\ \ldots\ \tilde{l}_N^J(k)\right]^T$, $\boldsymbol{s}(k) = \left[s_1^1(k)\ \ldots\ s_N^J(k)\right]^T$, $\mathbf{d}(k) = \left[d_1^1(k)\ \ldots\ d_N^J(k)\right]^T$, $\boldsymbol{\rho}(k) = \left[\rho_1^1(k)\ \ldots\ \rho_N^J(k)\right]^T$, $\bar{\boldsymbol{\Pi}} = \left[\bar{\Pi}_{1,1}\ \ldots\ \bar{\Pi}_{N,H}\right]^T$. The computational network dynamics can then be written as

$$\hat{\boldsymbol{d}}(\nu|k) = \min\left\{\hat{\tilde{\boldsymbol{l}}}(\nu|k) + \hat{A}(\nu|k)\hat{\boldsymbol{s}}(\nu|k),\ M\hat{\boldsymbol{\rho}}(\nu|k)\right\}, \tag{4}$$

$$\hat{\boldsymbol{l}}(\nu+1|k) = \hat{\tilde{\boldsymbol{l}}}(\nu|k) + \hat{A}(\nu|k)\hat{\boldsymbol{s}}(\nu|k) - \hat{\boldsymbol{d}}(\nu|k), \tag{5}$$

$$\mathbf{0} \leq \hat{\boldsymbol{s}}(\nu|k) \leq \mathbf{1}, \qquad \mathsf{S}\hat{\boldsymbol{s}}(\nu|k) \leq \mathbf{1}, \tag{6}$$

$$\hat{\tilde{\boldsymbol{l}}}(\nu|k) \geq \mathbf{0}, \qquad \mathsf{L}\hat{\tilde{\boldsymbol{l}}}(\nu|k) \leq \hat{\boldsymbol{l}}(\nu|k), \tag{7}$$

$$\mathbf{0} \leq \hat{\boldsymbol{\rho}}(\nu|k) \leq \mathbf{1}, \qquad \mathsf{B}_\rho\hat{\boldsymbol{\rho}}(\nu|k) \leq \bar{\boldsymbol{\Pi}}, \tag{8}$$

where the matrices $\hat{A}(\nu|k)$, M, S, L, and B_ρ can be constructed from (1)-(3) and the min operator in (4) is to be considered component-wise.

We consider a linear relationship between the average amount of hardware resources used by a server node over a certain interval and the average amount of power consumption of the node. Let $\beta_{i,h}$ represent the coefficient relating power consumption of the i^{th} node to the usage of the hardware resources in class h. The expected average power consumption of the i^{th} server node is

$$\hat{\mathsf{p}}_i(\nu|k) = \sum_{h=1}^{H}\sum_{j=1}^{J} \beta_{i,h}^j \hat{d}_i^j(\nu|k). \tag{9}$$

Eq. (9) can be rewritten as

$$\hat{\mathbf{p}}(\nu|k) = \mathsf{B}_d\hat{\boldsymbol{d}}(\nu|k), \tag{10}$$

where $\mathbf{p}(k) = [\mathsf{p}_1(k)\ \ldots\ \mathsf{p}_N(k)]^T$ and the matrix B_d depends on the coefficients $\{\beta_h^j\}$.

2.2 Thermal model

From the physical perspective, data center devices are modeled as nodes of a thermal network and they are divided into three classes: *server*, *CRAC*, and *environment*. Thermal server nodes represent the counterpart of the computational server nodes. Nodes in the CRAC class represent *computer room air conditioner* units and their related components. Nodes in the environment class represent support devices, e.g., uninterruptible power supplies (UPS), power distribution units (PDU), and network related components. Environment nodes are further divided into two groups. In the first group we include devices that consume power and for which meaningful inlet and outlet temperatures (later discussed) can be defined. The second class of environment nodes represents pure heat sources, e.g., the external weather.

Since this paper focuses on a controller for the whole data center, the effect of the environment nodes is considered negligible. For the sake of completeness, environment nodes will be included during the development of

the thermal model, but their effects will be thenceforth disregarded.

Thermal nodes are ordered. Indexes from 1 to N represent server nodes, indexes from $N+1$ up to $N+C$ represent CRAC nodes, indexes from $N+C+1$ up to $N+C+E_1$ represent the nodes in the first group of the environment class, and indexes from $N+C+E_1+1$ up to $N+C+E_1+E_2$ represents the nodes in the second group of the environment class. Let us define the sets $\mathcal{N}=\{1,\ldots,N\}$, $\mathcal{C}=\{N+1,\ldots,N+C\}$, $\mathcal{E}_1=\{N+C+1,\ldots,N+C+E_1\}$, and $\mathcal{E}_2=\{N+C+E_1+1,\ldots,N+C+E_1+E_2\}$. Let $\boldsymbol{x}$ be an $n\times 1$ vector and $\mathcal{I}$ be a nonempty, ordered set of indexes having values between 1 and n. With $\boldsymbol{x}_{\mathcal{I}}$ we denote the vector $[x_{i_1}\ \ldots\ x_{i_{|\mathcal{I}|}}]^T$, where i_j is the j^{th} element of $\mathcal{I}$ and $|\mathcal{I}|$ is the cardinality of the set $\mathcal{I}$. With $diag\{\boldsymbol{x}\}$ we denote the diagonal matrices having x_i as its i^{th} element along the main diagonal.

For every thermal node we define an output temperature, whereas only for those nodes having index in $\mathcal{N}\cup\mathcal{C}\cup\mathcal{E}_1$ we define an input temperature. The output temperature of the i^{th} node at time t is denoted by $T_{\text{out},i}(t)$ and it represents the average temperature of the air flowing out of the node at time t. The input temperature of the i^{th} node at time t is denoted by $T_{\text{in},i}(k)$ and it represents the average temperature of the air flowing into the node at time t.

As discussed in the work of Tang et al. (2006), a linear relationship is considered between the output temperature and the input temperature of every node. Let $\psi_{i,j}$ denote the thermal coupling between the input temperature of the i^{th} thermal node and the output temperature of the j^{th} thermal node. The coefficients $\{\psi_{i,j}\}$ are non-negative and $\sum_j \psi_{i,j}=1$ for all $i\in\mathcal{N}\cup\mathcal{C}\cup\mathcal{E}_1$. For all $i,j\in\mathcal{N}\cup\mathcal{C}\cup\mathcal{E}_1$, we define the matrix $[\Psi]_{i,j}=\psi_{i,j}$ and for $i\in\mathcal{N}\cup\mathcal{C}\cup\mathcal{E}_1$ and $j\in\mathcal{E}_2$, we define the matrix $[\Psi_{E_2}]_{i,j}=\psi_{i,j}$. The relationship between the input and the output temperature of every node can now be written as

$$\hat{\boldsymbol{T}}_{\text{in}}(\tau|t)=\Psi\hat{\boldsymbol{T}}_{\text{out}}(\tau|t)+\Psi_{E_2}\hat{\boldsymbol{T}}_{\text{out},E_2}(\tau|t), \tag{11}$$

where we defined the vectors

$$\boldsymbol{T}_{\text{in}}(t)=[T_{\text{in},1}(t)\ \ldots\ T_{\text{in},N+C+E_1}(t)]^T$$
$$\boldsymbol{T}_{\text{out}}(t)=[T_{\text{out},1}(t)\ \ldots\ T_{\text{out},N+C+E_1}(t)]^T$$
$$\boldsymbol{T}_{\text{out},E_2}(t)=[T_{\text{out},N+C+E1+1}(t)\ldots T_{\text{out},N+C+E_1+E_2}(t)]^T.$$

The thermal constraints are stated in terms of the node inlet temperatures and can be written as

$$\hat{\boldsymbol{T}}_{\text{in}}(\tau|t)\leq\overline{\boldsymbol{T}_{\text{in}}}. \tag{12}$$

Server nodes. A linear model is used to predict the evolution of the server node output temperatures

$$\dot{\hat{T}}_{\text{out},i}(\tau|t)=-\mathsf{k}_i(\hat{T}_{\text{out},i}(\tau|t)-\hat{T}_{\text{in},i}(\tau|t))+\mathsf{c}_i\hat{\mathsf{p}}_i(\tau|t), \tag{13}$$

where $\hat{\mathsf{p}}_i(\tau|t)$ is the expected power consumption of the i^{th} server node at time τ, k_i is the inverse of the time constant of the node, and c_i is the coefficient that transforms electric power into temperature variations.

CRAC nodes. We assume that the evolution of the air flowing out from a CRAC unit can be modeled as

$$\begin{aligned}\dot{\hat{T}}_{\text{out},i}(\tau|t)=&-\mathsf{k}_i\hat{T}_{\text{out},i}(\tau|t)+\\&\mathsf{k}_i\min\left\{\hat{T}_{\text{in},i}(\tau|t),\hat{T}_{\text{ref},i-N}(\tau|t)\right\}.\end{aligned} \tag{14}$$

The reference temperature of the i^{th} CRAC unit takes values in the interval $[\underline{T_{\text{ref}}}_{i-N},\overline{T_{\text{ref}}}_{i-N}]$. The model for the power consumption of the i^{th} CRAC unit is derived from the model described in the work of Moore et al. (2005). Let f_i be the rate at which air flows into the i^{th} CRAC unit and c_p be the specific heat of the air at standard pressure. The expected power consumption of the i^{th} CRAC node at time τ is

$$\mathsf{p}_i(\tau|t)=\begin{cases}\mathsf{c}_p\mathsf{f}_i\dfrac{\hat{T}_{\text{in},i}(\tau|t)-\hat{T}_{\text{out},i}(\tau|t)}{COP_i(\tau|t)}\\ \qquad\qquad\text{if } \hat{T}_{\text{in},i}(\tau|t)\geq\hat{T}_{\text{out},i}(\tau|t)\\ 0\qquad\qquad\text{otherwise}\end{cases}, \tag{15}$$

where $COP_i(\tau|t)$ is the expected value of the *coefficient of performance* (COP) of the i^{th} CRAC unit at time τ. The COP of a conditioner unit is function of its outlet air temperature.

Environment nodes. A linear model similar to the one used for thermal server nodes is used to describe the evolution of the output temperature of the environment nodes of the first group. For all $i\in\mathcal{E}_1$, we can write

$$\dot{\hat{T}}_{\text{out},i}(\tau|t)=-\mathsf{k}_i\hat{T}_{\text{out},i}(\tau|t)+\mathsf{k}_i\hat{T}_{\text{in},i}(\tau|t)+\mathsf{c}_i\hat{\mathsf{p}}_i(\tau|t), \tag{16}$$

and the average power consumption is considered an exogenous uncontrollable input. The output temperature of nodes in the second group of environment nodes is considered an exogenous time varying input to the data center.

Discrete-time model. Let us define the vectors $\boldsymbol{T}_{\text{ref}}(t)=[T_{\text{ref},1}(t)\ \ldots\ T_{\text{ref},C}(t)]^T$, $\mathbf{k}=[\mathsf{k}_1\ \ldots\ \mathsf{k}_{N+C+E_1}]^T$, $\mathbf{p}(t)=[\mathsf{p}_1(t)\ \ldots\mathsf{p}_{N+C+E_1}(t)]^T$. If for all $i\in\mathcal{C}$ $T_{\text{ref},i}(t)$ is constrained to be less than or equal to $T_{\text{in},i}(t)$, then the evolution of the thermal network can be written as

$$\begin{aligned}&\dot{\hat{\boldsymbol{T}}}_{\text{out}}(\tau|t)=A_{T,C}\hat{\boldsymbol{T}}_{\text{out}}(\tau|t)+\\&B_{T,C}\left[\hat{\mathbf{p}}_{\mathcal{N}}(\tau|t)^T\ \hat{\boldsymbol{T}}_{\text{ref}}(\tau|t)^T\ \hat{\mathbf{p}}_{\mathcal{E}_1}(\tau|t)^T\ \hat{\boldsymbol{T}}_{\text{out},E_2}(\tau|t)^T\right]^T,\end{aligned} \tag{17}$$

where the values of the matrices $A_{T,C}$ and $B_{T,C}$ can be derived from (11)-(16).

Henceforth we focus on a the case where no environment nodes are included in the thermal model. A discrete-time version of (17) can be written as

$$\begin{aligned}\hat{\boldsymbol{T}}_{\text{out}}(\nu+1|k)=&A_{T,D}\hat{\boldsymbol{T}}_{\text{out}}(\nu|k)+\\&B_{T,D}\left[\hat{\mathbf{p}}_{\mathcal{N}}(\nu|k)^T\,\hat{\boldsymbol{T}}_{\text{ref}}(\nu|k)^T\right]^T.\end{aligned} \tag{18}$$

The dynamic system in (18) is marginally stable and the matrices $A_{T,D}$ and $B_{T,D}$ have all positive elements.

2.3 Cost function

Three metrics are included in the cost function: the expected average QoS cost (19), the cost of the chosen control action (20), and the expected cost of powering the data center (21) denoted $c_{Qos}(k)$, $c_{l,s}(k)$, and $c_e(k)$ respectively. Concerning the expected QoS cost, we consider the approach discussed in the work of Zhu et al. (2008). Let $c_h^{j,\star}$ denote the optimal ratio between the amount of resources available and the amount of resources necessary in class h

for jobs in class j. The expected QoS cost can be written as

$$\hat{c}_{Qos}(\nu|k) = \sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{i=1}^{N} \left((\hat{\tilde{l}}_i^j(\nu|k) + \hat{a}_i^j(\nu|k))b_h^j c_h^{j,\star} - \hat{\rho}_i^j(\nu|k)\tilde{b}_{i,h}^j\right)^2. \quad (19)$$

To favor control actions that reduce the number of migrating VMs and also the number of dropped jobs, we define the cost for the control action

$$\hat{c}_{l,s}(\nu|k) = \sum_{i=1}^{N}\sum_{j=1}^{J}(\hat{l}_i^j(\nu|k) - \hat{\tilde{l}}_i^j(\nu|k))^2 - (\mathsf{c}_{s,i}^j s_i^j(\nu|k))^2. \quad (20)$$

Let $\bar{\mathsf{p}}(k)$ be the threshold on the average power consumption during the k^{th} interval which allows the data center to pay a reduced cost of the electricity. Define $\mathsf{p}(k)$ as the total average data center power consumption during the k^{th} interval. Let $\alpha_e(k)$ be the cost of electricity up to $\bar{\mathsf{p}}(k)$ and $\beta_e(k)$ be the cost of electricity over $\bar{\mathsf{p}}(k)$. We can write

$$\hat{c}_e(\nu|k) = \begin{cases} \hat{\alpha_e}(\nu|k)\hat{\mathsf{p}}(\nu|k) & \hat{\mathsf{p}}(\nu|k) \le \hat{\bar{\mathsf{p}}}(\nu|k) \\ \hat{\alpha_e}(\nu|k)\hat{\bar{\mathsf{p}}}(\nu|k) + \hat{\beta_e}(\nu|k)\left(\hat{\mathsf{p}}(\nu|k) - \hat{\bar{\mathsf{p}}}(\nu|k)\right) & \\ \qquad \hat{\mathsf{p}}(\nu|k) > \hat{\bar{\mathsf{p}}}(\nu|k). & \end{cases} \quad (21)$$

The total cost at time ν given the estimation at time k is

$$J(\nu|k) = \mathsf{c}_Q \hat{c}_{Qos}(\nu|k) + \hat{c}_e(\nu|k) + \mathsf{c}_{\delta,s}\hat{c}_{l,s}(\nu|k). \quad (22)$$

2.4 Optimal control problem

Let $\mathcal{T} \in \mathbb{N}$ be the horizon for the optimal control problem, and define the sets $\mathcal{R} = \{\hat{\boldsymbol{\rho}}(k|k), \ldots, \hat{\boldsymbol{\rho}}(k+\mathcal{T}|k)\}$, $\mathcal{T}_{\text{ref}} = \{\hat{\boldsymbol{T}}_{\text{ref}}(k|k), \ldots, \hat{\boldsymbol{T}}_{\text{ref}}(k+\mathcal{T}|k)\}$, $\mathcal{S} = \{\hat{\boldsymbol{s}}(k|k), \ldots, \hat{\boldsymbol{s}}(k+\mathcal{T}|k)\}$, $\tilde{\mathcal{L}} = \{\hat{\tilde{\boldsymbol{l}}}(k|k), \ldots, \hat{\tilde{\boldsymbol{l}}}(k+\mathcal{T}|k)\}$. At every time k, the optimal control problem that the data center controller has to solve is

$$\begin{aligned} \min_{\mathcal{R},\mathcal{T}_{\text{ref}},\mathcal{S},\tilde{\mathcal{L}}} & \sum_{\nu=k}^{\mathcal{T}+k} J(\nu|k) \\ \textbf{s.t.} & \\ & \text{for all } \nu = k, \ldots, k+\mathcal{T} \\ & (4)-(8), (18) \\ & \hat{\boldsymbol{p}}_{\mathcal{N}}(\nu|k) = \mathsf{B}_d \hat{\boldsymbol{d}}(\nu|k) \\ & \hat{\boldsymbol{T}}_{\text{in}}(\nu+1|k) \le \overline{\boldsymbol{T}_{\text{in}}} \\ & \underline{\boldsymbol{T}_{\text{ref}}} \le \hat{\boldsymbol{T}}_{\text{ref}}(\nu|k) \le \overline{\boldsymbol{T}_{\text{ref}}} \\ & \hat{\boldsymbol{T}}_{\text{out}}(k|k) = \boldsymbol{T}_{\text{out}}(k) \\ & \hat{\boldsymbol{l}}(k|k) = \boldsymbol{l}(k). \end{aligned} \quad (23)$$

Due to the positivity of the elements in the matrices Ψ, $A_{D,T}$, and $B_{D,T}$, the optimization problem in (23) is feasible if for all $\boldsymbol{x}$ such that $\Psi\boldsymbol{x} = \overline{\boldsymbol{T}_{\text{in}}}$

$$\Psi A_{D,T}\boldsymbol{x} + \Psi B_{D,T}\left[\mathbf{0}^T \ \underline{\hat{\boldsymbol{T}}_{\text{ref}}}^T\right]^T \le \overline{\boldsymbol{T}_{\text{in}}}. \quad (24)$$

Sub-optimal approach. We consider the following additional constraint

$$M\hat{\boldsymbol{\rho}}(\nu|k) \ge \hat{\tilde{\boldsymbol{l}}}(\nu|k) + A(k)\hat{\boldsymbol{s}}(\nu|k). \quad (25)$$

When (25) is considered, then the evolution of the computational sub-system can be disregarded and the overall dynamic of the data center is completely described by the thermal sub-system. In the rest of the paper therefore, we apply (25).

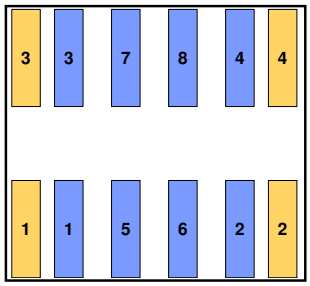

Fig. 1. Data center layout: 8 server nodes (blue) and 4 CRAC nodes (orange).

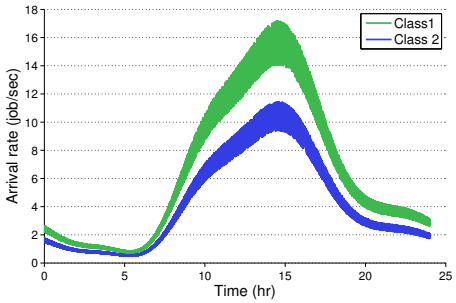

Fig. 2. Job arrival rates at the data center.

3. SIMULATION

We consider the data center layout depicted in Fig. 1. Server nodes represent a collection of 3 racks, each composed of 42 servers. Server nodes 5-8 are slightly more efficient than the other server nodes, but they also have large self-correlating temperature coefficients: $\psi_{i,i} \sim 0.5$ for $i = 5, \ldots, 8$. All server nodes receive cool air from at least one CRAC node, but server nodes 1-4 are more efficiently cooled than the other nodes. The CRAC nodes are identical. Each CRAC node is strongly coupled with the most nearby server node.

Two job classes and two hardware classes are considered. Job arrivals are subject to a random noise uniformly distributed having zero mean and a variance proportional to the mean arrival rate.

Simulations were developed using the TomSym language and KNITRO was used as numerical solver.[2] The time-step for the simulation is 30 s and the controller closes the loop every 10 min. The optimization problem is formulated using a horizon of 6 steps (one hour). The time constants of the server nodes are all equal to 3 min. The time constant for the CRAC node is approximately 1 min.

We call the controller solving the relaxation of (23), the *coordinated* controller. The control strategy obtained by the coordinated controller is compared against a controller which does not consider the coupling between the computational and the thermal part of the data center. We call the latter controller *uncoordinated*.

At every time k the uncoordinated controller solves two optimization problems. In the first step, the uncoordinated controller chooses a control action that minimizes the sum of the expected QoS cost, the server powering cost, and the cost of the chosen control action

[2] http://tomsym.com/ and http://www.ziena.com/knitro.html .

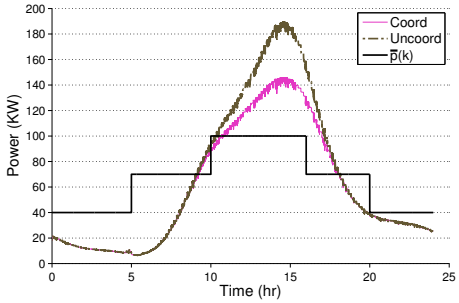

Fig. 3. Total data center power consumption.

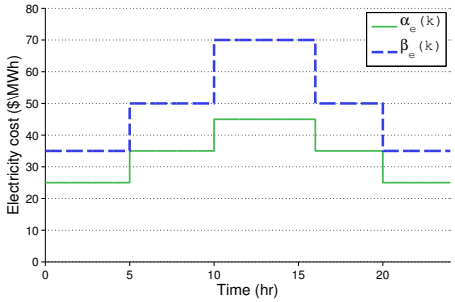

Fig. 4. Electricity cost over time.

$$\begin{aligned} &\min_{\mathcal{R},\mathcal{S},\tilde{\mathcal{L}}} \sum_{\nu=k}^{\mathcal{T}+k} J_{U,1}(\nu|k) \\ &\textbf{s.t.} \\ &\quad \text{for all } \nu = k,\dots,k+\mathcal{T} \\ &\quad (4)-(8) \\ &\quad \hat{\boldsymbol{p}}_{\mathcal{N}}(\nu|k) = \mathsf{B}_d \hat{\boldsymbol{d}}(\nu|k) \\ &\quad \hat{\boldsymbol{l}}(k|k) = \boldsymbol{l}(k), \end{aligned} \tag{26}$$

where

$$\begin{aligned} J_{U,1}(\nu|k) = \mathsf{c}_Q \hat{c}_{Qos}(\nu|k) + \mathsf{c}_{\delta,s}\hat{c}_{l,s}(\nu|k) + \\ \hat{\alpha}_e(\nu|k)\hat{\mathsf{p}}_{\mathcal{N}}(\nu|k). \end{aligned} \tag{27}$$

In (27) $\hat{\mathsf{p}}_{\mathcal{N}}(\nu|k)$ represents the sum of the expected average power consumption of server nodes.

In the second step, the uncoordinated controller, based on the solution obtained at the first step, solves the following optimization problem to select the best vector $\boldsymbol{T}_{\text{ref}}(k)$ that minimizes the cost of powering the CRAC nodes and that also enforces the thermal constraints

$$\begin{aligned} &\min_{\mathcal{T}_{\text{ref}}} \sum_{\nu=k}^{\mathcal{T}+k} \hat{\alpha}_e(\nu|k)\hat{\mathsf{p}}_{\mathcal{C}}(\nu|k) \\ &\textbf{s.t.} \\ &\quad \text{for all } \nu = k,\dots,k+\mathcal{T} \\ &\quad (18) \\ &\quad \hat{\boldsymbol{T}}_{\text{in}}(\nu+1|k) \le \overline{\boldsymbol{T}_{\text{in}}} \\ &\quad \underline{\boldsymbol{T}_{\text{ref}}} \le \hat{\boldsymbol{T}}_{\text{ref}}(\nu|k) \le \overline{\boldsymbol{T}_{\text{ref}}} \\ &\quad \hat{\boldsymbol{T}}_{\text{out}}(k|k) = \boldsymbol{T}_{\text{out}}(k). \end{aligned} \tag{28}$$

The variable $\hat{\mathsf{p}}_{\mathcal{C}}(\nu|k)$ represents the sum of the expected average power consumption of CRAC nodes. The uncoordinated controller cannot compute the expected cost of powering the data center and hence it cannot leverage on the SLA with the grid.

In order to create a fair comparison among the actions chosen by the coordinated and the uncoordinated algorithms, we consider a high cost for dropping jobs. In this case, the two algorithms are forced to processes almost every incoming job. In the simulation, the ratio between the number of dropped jobs and the number of jobs arrived at the data center during a control actions was less than $6.8{\cdot}10^{-5}$ for both of the controllers and no jobs were lost due to the migration actions.

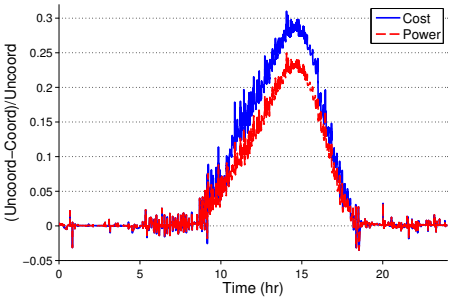

Fig. 5. Relative difference of the cost of powering and of the total power consumption.

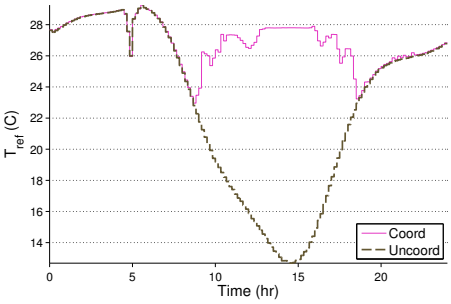

Fig. 6. Average reference temperature.

Figure 2 shows the arrival rate for jobs in class 1 and 2 at the data center. Job arrival rates represent a scaled version of the request rate arrived to an EPA server on Aug. 30^{th} 1995.[3] Figure 3 shows the total data center power consumption obtained by the coordinated and by the uncoordinated controller. The power threshold value ($\bar{\mathsf{p}}(k)$) is also shown in Fig. 3. The cost of electricity over time is shown in Fig. 4. The relative ratio between cost of powering the data center for the uncoordinated and the coordinated controller is shown in Fig. 5. The relative ratio between the power consumption of the data center when the two controllers are used is also shown in Fig. 5.

Both controllers kept, in the average, the optimal ratio between the amount of hardware resources allocated and the amount of hardware resources required. The oscillations around the mean value had a standard deviation of 0.05. We expect that lower level controllers, not implemented in this simulation, can reduce the oscillation around the optimal value since they control server nodes at a much faster rate.

Figure 6 shows the average CRAC reference temperature for the coordinated and the uncoordinated controller case. Both controllers over-cool the servers just before the variation of the electricity cost at time 5 hr and they are able to reduce almost to zero the usage of CRAC unit for the following 30 min. Figure 7 shows the total power consumption the of server and of the CRAC nodes obtained by the coordinated and the uncoordinated controllers. The coordinated controller, even though it obtains slightly higher values for the total server power consumption, is able to maintain a higher level of cooling efficiency. Therefore, it

[3] Source: The Internet Traffic Archive http://ita.ee.lbl.gov/ .

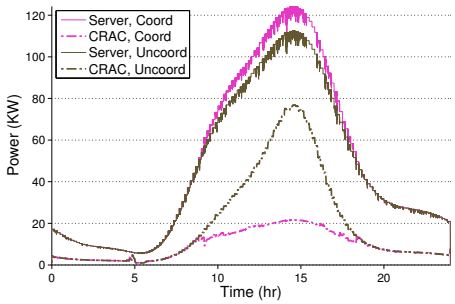

Fig. 7. Power consumption values of server nodes and CRAC nodes.

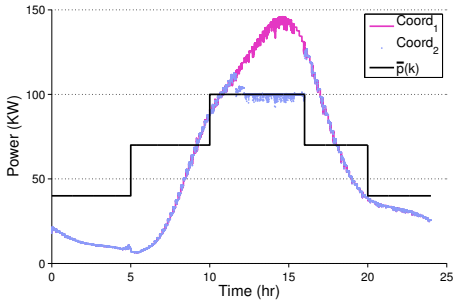

Fig. 8. Total data center power consumption.

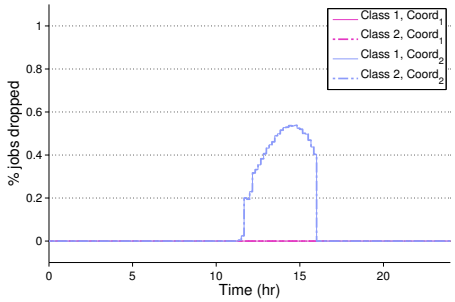

Fig. 9. Relative percentage of dropped jobs.

is able to largely reduce the power consumption of the CRAC units with respect to the uncoordinated controller.

From Fig. 6 and 7 it can also be observed how the actions of the two controllers lead to the same server and CRAC power consumption as well as to the same average reference temperatures until the average arrival rate is lower than a certain threshold, e.g., just before time 10 hr and a little before time 20 hr. This particular behavior suggests that the uncoordinated policy is almost optimal as long as the average value of the arrival rate is reduced.

Figure 8 shows the total data center power consumption when two coordinated controllers are used. The first coordinated controller is the same as the one derived in the previous simulation and has a large cost for dropping jobs. The second coordinated controller is derived using a reduced cost for dropping jobs. As shown in Fig. 8 the two control actions differ only when the cost of the current is over a certain value. Figure 9 shows the average percentage of jobs dropped over the time by the two controllers.

In the simulation both of the controllers were able to enforce the inlet temperature constraints at discrete time, but not during the complete evolution. This result was expected since the constraints were formulated only for the discrete-time evolution.

4. DISCUSSION

A model that includes the computational and the physical characteristics as well as their interaction is described in this paper. Two cases are considered in the simulation section. In the first one, the proposed controller is forced to process all of the incoming jobs, whereas in the latter, the controller is allowed to drop jobs. The simulations provide an example where the total power consumption of the data center changes abruptly based on the variations of the job arrival rate and of the electricity cost. In particular, the cost of electricity and the threshold power value set by the grid do not always force a data center to cap its own power consumption. The average job arrival rate also plays a relevant role. Further research is necessary to understand how an SLA with data center mangers should be stipulated in order to allow the data center to become an efficient smart node of the power grid.

The primary goal of a data center control strategy is to minimize a given cost function, regardless of the stability of the overall system. This leads to the question of how controllers at different levels should operate so as to minimize the overall cost function. Some preliminary results toward this goal were discussed in [Parolini et al. (2010a)]. Other relevant work in this sense can be found, among others, in [Rawlings and Amri (2009)].

REFERENCES

Fan, X., Weber, W.D., and Barroso, L.A. (2007). Power provisioning for a warehouse-sized computer. In *International Symposium on Computer Architecture*.

Hamilton, J. (2008). Cost of power in large-scale data centers. `http://perspectives.mvdirona.com`.

Moore, J., Chase, J., Ranganathan, P., and Sharma, R. (2005). Making scheduling "Cool": temperature-aware workload placement in data centers. In *USENIX Annual Technical Conference*.

Parolini, L., Garone, E., Sinopoli, B., and Krogh, B.H. (2010a). Cyber-physical approach to data center modeling and control.

Parolini, L., Tolia, N., Sinopoli, B., and Krogh, B.H. (2010b). A cyber-physical systems approach to energy management in data centers. In *First international conference on cyber-physical systems*.

Rawlings, J.B. and Amri, R. (2009). *Nonlinear Model Predictive Control*, chapter Optimizing Process Economic Performance Using Model Predictive Control, 119–138. Springer Berlin / Heidelberg.

Sharma, R.K., Bash, C.E., Patel, C.D., Friedrich, R.J., and Chase, J.S. (2005). Balance of power: Dynamic thermal management for internet data centers. *IEEE Internet Computing*, 9, 42–49.

Tang, Q., Mukherjee, T., Gupta, S.K.S., and Cayton, P. (2006). Sensor-based fast thermal evaluation model for energy efficient high-performance data centers. In *Intelligent Sensing and Information Processing*.

U.S. Environmental Protection Agency (2007). Report to congress on server and data center energy efficiency. Technical report, ENERGY STAR Program.

Zhu, X., Young, D., Watson, B., Wang, Z., Rolia, J., Singhal, S., McKee, B., Hyser, C., Gmach, D., Gardner, R., Christian, T., and Cherkasova, L. (2008). 1000 islands: Integrated capacity and workload management for the next generation data center. In *International Conference on Autonomic Computing*.