Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Using Model Predictive Control in Data Centers for Dynamic Server Provisioning ⋆

Qiu Fang * **Jun Wang** *,1 **Han Zhu** * **Qi Gong** **

* *Department of Control Science and Engineering, Tongji University, Shanghai 201804 P. R. China (e-mail: {08fangqiu—junwang—13sdaqzh}@tongji.edu.cn)*
** *Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064 USA (e-mail: qigong@soe.ucsc.edu)*

Abstract: The ever-increasing energy consumption of data centers is becoming a crucial problem nowadays. This paper discusses the benefits and challenges of coordinating dynamic server provisioning and thermal dynamics for energy control of data centers. A model predictive control approach, considering the computational and thermal characteristics of a data center and their interactions, together with dynamic server provisioning is used to optimize the energy consumption of the total data center. Quality of service and thermal constraints are enforced in the optimization process. Simulation results demonstrate that the proposed control approach can lead to a significant energy saving for a data center.

Keywords: Efficiency enhancement; Energy control; Optimization; Predictive control.

1. INTRODUCTION

The energy consumption of high performance data centers increased significantly with high density clusters and server farms. *The global census on data center trends* [2] showed that data center power requirements grew up to globally 38GW (gigawatts) in 2012, about 63% growth from 24GW in 2011. Data centers worldwide are estimated to account for over 2% of global greenhouse gas emissions in 2011 [3]. Data centers are main consist of information technology (IT) system and cooling technology (CT) system. The IT system cost more than half of the total power consumption of a data center. *The Uptime Institute 2012 Data Center Industry Survey* [4] shows the average PUE of their respondents largest data centers is between 1.8 and 1.89, where PUE is the power of usage effectiveness, which is used as a measure of data center efficiency in industry. The PUE is defined as the ratio of the total facility power consumption and the power consumption of the IT. Regarding the huge energy consumption and high PUE of data centers, it is essential to improve the energy efficiency of data center operations.

In previous works, Parolini et al. (2012) and Parolini et al. (2011) developed a cyber-physical system model at data center level, and applied an model predictive control (MPC) method to improve data center energy efficiency and discussed the potential of energy saving. Tang et al. (2006) and Tang (2008) developed a fast temperature evaluation model of data center thermal dynamics, which could help implement real-time control in data center while considering the control of CRAC units. Chen et al. (2008) discussed several server provisioning algorithms to dynamic turn servers on/off for the purpose of saving energy. Doyle et al. (2003) and Chase et al. (2001) studied several dynamic provisioning and load patching algorithms. In the research of Parolini et al. (2012) and Parolini et al. (2011) details of server level dynamics was neglected, and models of servers were regarded as simple linear models at data center level. Their work also lacks discussion of the quality of service (QoS), while the QoS is an important aspect in the control of data center. Dynamic server provisioning and load patching algorithms just took computational dynamics into consideration, regardless of the location of servers which related to cooling efficiency. An unreasonable active server distribution may easily produce hot spots in data center. If implement dynamic server provisioning while considering thermal dynamics could help prevent hot spots and achieve significant energy efficiency in both IT and CT system.

This paper considered a case where a data center do not cooling efficiently. In the precondition of satisfying QoS requirements, we focused on the energy saving offered by implementing dynamic server provisioning while considering data center thermal dynamics. As IT system consuming more than half of the total power consumption, dynamic server provisioning could largely decrease the power consumption of IT system by closing idle servers on demands. At the same time different areas of data center have different cooling efficiency, dynamically allocating tasks to server considering their thermal dynamics could achieve optimal efficiency. To maximize profit, Several MPC approaches considering different situations

⋆ This work was supported by the International Science and Technology Cooperation Program of China under Grant No. 2011DFG13020.

[1] Corresponding author, e-mail: junwang@tongji.edu.cn.

[2] `http://www.computerweekly.com/news/2240164589/Datacentre-power-demand-grew-63-in-2012-Global-datacentre-census`

[3] `http://www.greenpeace.org/international/en/publications/reports/How-dirty-is-your-data`

[4] `http://www.uptimeinstitute.com/2012-survey-results`

are used in data centers to compare the energy efficiency performance. By optimizing active server numbers and reference temperatures of computer room air conditioner (CRAC) units, the MPC controllers achieved a significant improvement of energy efficiency.

The remainder of the paper is organized as follows. Section 2 presents a data center model. Section 3 considers control strategies used in this paper. Section 4 presents simulation results. Section 5 makes concluding remarks.

2. PROBLEM STATEMENTS

A data center mainly consists of the IT system and the CT system. In the IT part, racks of servers are grouped into zones. In the CT part several CRAC units are used to cool down the data center. In this paper, we consider connection service as the basic service of the data center, each server in zones runs only one connection service application.

2.1 Zone level model

As presented in Chen et al. (2008), in a front door architecture for connection intensive applications, every login request sent to the zone from end users will reach a dispatch server first. The dispatch server picks a connection server and returns its IP address to the client. Then the client directly connects to the connection server. The connection server authenticates the user and if succeeded, a live TCP connection will be maintained between the client and the connection server until the client logs off. The TCP connection is usually used to update user status (e.g. on-line, busy, off-line etc.) and to redirect further activities such as chatting and multimedia conferencing to other back-end servers.

System Dynamics At an application level, each connection server has to enforce two major constraints: the maximum login rate and the maximum number of connections it can host. The login rate L is defined as the number of new connection requests which sent to a connection server in a second. A limit on login rate $L_{\max}$ is used to protect the server. For the consideration of memory constraints and fault tolerance concerns, a limit $N_{\max}$ is considered on the total number of connections for each connection server.

This paper assumes that a zone consists of $H_{\max}$ connection servers. Let $H_i(t)$ denote the number of available servers in Zone i, $N_\delta(t)$ the number of connections on Server δ, and $L_\delta(t)$ and $D_\delta(t)$ the number of login rate and departure rate respectively. The dynamics of the δth server can be modeled as

$$\dot{N}_\delta(t) = L_\delta(t) - D_\delta(t), \qquad \delta = 1, \cdots, H_i(t) \tag{1}$$

This differential equation represents the relationship between the login rate $L_\delta(t)$ and the number of connections $N_\delta(t)$. The number of departures $D_\delta(t)$ is usually a part of $N_\delta(t)$, which varies a lot with time. The login rate $L_\delta(t)$ dispatched to one of available servers in Zone i is a part of total login rate $L_{Z,i}(t)$. A dispatch algorithm can be expressed as

$$L_\delta(t) = L_{Z,i}(t) \cdot p_\delta(t), \qquad \delta = 1, \cdots, H_i(t) \tag{2}$$

where $p_\delta(t)$ is the fraction of the total login requests assigned to Server δ, $0 \le p_\delta(t) \le 1$, $\sum_{\delta=1}^{H_i(t)} p_\delta(t) = 1$ and $H_i(t)$ the number of available servers in Zone i.

Performance model The performance model of a connection server is derived from the model described in the work of Chen et al. (2008). The key variables affecting CPU usage and power of connection Server δ are login rates $L_\delta(t)$ and the number of active connections $N_\delta(t)$. The linear model is as below

$$\hat{U}_\delta(t) = 2.84 \times 10^{-4} \cdot N_\delta(t) + 0.549 \cdot L_\delta(t) - 0.820 \tag{3}$$

where $\hat{U}_\delta(t)$ denotes the CPU utilization percentage. The power consumption of a connection server increases almost linearly with CPU utilization, while the idle servers consume up to almost 60% of the peak power. The power consumption of Server δ is modeled as

$$P_{s,\delta}(t) = P_{s,0} + \eta \cdot U_\delta(t) \tag{4}$$

where $P_{s,0}$ is the power consumption of idle servers, and η is a positive coefficient.

2.2 Data-center level model

At data-center level, As discussed in the work of Parolini et al. (2011), the thermal properties of IT system are managed as groups of components (racks of servers). Servers are aggregated into several zones, and each zone has both computational and thermal dynamics. The IT system processes clients' requests, and maintains active connections, it consumes power and generates heat at the meantime. The CT system removes heat from zones by consuming power.

In this section, the data-center level dynamics model formulated from the computational and thermal properties and their relationship. The data-center level model studied in this paper has N computation nodes and M thermal nodes, i.e. N zones and $M - N$ CRAC units in this model. Just like the data center considered in the following simulation part as Fig. 1, there is 4 computational nodes and 6 thermal nodes. In the following sections $i = 1, \cdots, N$ represent the number of each zone, and $i = N+1, \cdots, M$ for each CRAC unit. The MPC method is utilized to

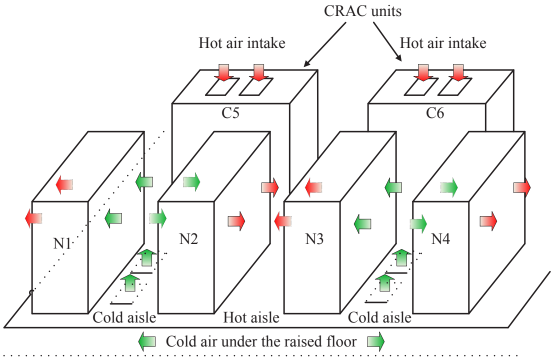

Fig. 1. Data center layout: 4 server zones ($N1 - N4$) and 2 CRAC units ($C5 - C6$)

save the potential energy by dynamically shutting down idle servers and changing the operation conditions of each component of the data center.

Computational model Zones are modeled as a single computational node. Zone level model shows that Zone i has a total login rate $L_{Z,i}(t)$, and a limit on it $L_{Z,\max}(t)$. The limit $L_{Z,\max}$ is affected by the number of available servers $H_i(t)$ ($L_{Z,max} = H_i(t) \cdot L_{\max}$). For a zone having $H_{\max}$ servers in all, $H_i(t) \leq H_{\max}$. Zone i also has a total connection departure rate $D_{Z,i}$ and a total number of active connections $N_{Z,i}(t)$. Given a data center with N computation nodes and M thermal nodes, the dynamics of Zone i can be expressed as the following differential equation

$$\dot{N}_{Z,i}(t) = L_{Z,i}(t) - D_{Z,i}(t), \qquad i = 1, \cdots, N \tag{5}$$

The total login rate $L_{Z,i}(t)$ of Zone i is dispatched by the data-center level controller. The dispatch algorithm is as below

$$L_{Z,i}(t) = L_{DC}(t) \cdot q_i(t), \qquad i = 1, \cdots, N \tag{6}$$

where $L_{DC}(t)$ is the total client requests sent to the data center, $q_i(t)$ the fraction of the total login requests assigned to Zone i, $0 \leq q_i(t) \leq 1$, $\sum_{i=1}^{N} q_i(t) = 1$.

The model of requests execution developed above is sufficient for our purpose. It can also be extended to include more components, such as different workload dispatch policies, job classes, hardware requirements and interactions among different workload classes.

Thermal model From a thermal perspective, zones, CRAC units, and other support devices are modeled as thermal nodes. Since this section focuses on the data-center level, the slight effect of support devices on the thermal environment is neglected. This paper has M thermal nodes in the data center. For each thermal node we define an output temperature and an input temperature. Tang et al. (2006) presents the input temperature of Node i has a relationship with the rate of the amount of heat received from the other thermal nodes, and the input temperature is denoted as $T_{in,i}(t)$, the rate of the amount of heat brought into Node i is denoted as $Q_{in,i}(t)$. The relationship between $T_{in,i}(t)$ and $Q_{in,i}(t)$ is given by

$$Q_{in,i}(t) = \rho \cdot f_i \cdot C_p \cdot T_{in,i}(t), \qquad i = 1, \cdots, M \tag{7}$$

where ρ is the air density, f_i the flow rate of Node i, and C_p the specific heat of air. The output temperature of Node i is defined as $T_{out,i}(t)$, and the rate of the amount of heat taken away from the node is denoted as $Q_{out,i}(t)$. The relationship between $T_{out,i}(t)$ and $Q_{out,i}(t)$ is the same as Equation (7). The variation between $Q_{in,i}(t)$ and $Q_{out,i}(t)$ has a relationship with the power consumption of Node i

$$Q_{out,i}(t) - Q_{in,i}(t) = \lambda_i \cdot P_i(t), \qquad i = 1, \cdots, M \tag{8}$$

Where λ_i is the coefficient of the power consumption of Node i converted to the heat. For server nodes, $\lambda_i \doteq 1$, $i = 1, \cdots, N$, assume that almost all of the power consumption converted to the heat. CRAC nodes remove heat from air flow and λ_i is a negative variable, $\lambda_i < 0$, $i = N + 1, \cdots, M$.

The rate of the amount of input heat $Q_{in,i}(t)$ carried by inlet air flow is a mixture of supplied cold air flow from CRAC nodes and recirculated hot air from other server nodes, so there is

$$Q_{in,i}(t) = \sum_{j=1}^{M} \phi_{j,i} \cdot Q_{out,j}(t), \qquad i = 1, \cdots, M \tag{9}$$

where the coefficient $\phi_{j,i}$ is the percentage of the heat flow from node j to Node i. The matrix $\Phi = [\phi_{j,i}]_{M \times M}$ is defined as the cross-interference among all thermal nodes, $\phi_{i,j}$ is nonnegative and $\sum_{j=1}^{M} \phi_{i,j} = 1$, $i = 1, ..., M$. Considering (7), (9) and substituting $\rho f_i C_p$ with K_i, for Node i we have

$$T_{in,i} = \frac{\sum_{j=1}^{M} \phi_{j,i} \cdot K_j \cdot T_{out,j}}{K_i}, \qquad i = 1, \cdots, M \tag{10}$$

Parolini et al. (2008) demonstrates that the revolution of the outlet temperature of a server node approximates a liner-invariant system, the revolution of the output temperature of the thermal nodes for zones is modeled by

$$\dot{T}_{out,i}(t) = -\alpha_i \cdot T_{out,i}(t) + \alpha_i \cdot T_{in,i}(t) + c_i \cdot P_i(t) \tag{11}$$

where $i = 1, \cdots, N$, and $1/\alpha_i$ is the time constant of the temperature of Node i, c_i is the coefficient that maps power consumption into output temperature variation and $P_i(t)$ is the power consumption of Node i. Considering the zone level model Equations (3),(4) and server On-Off state control. The power consumption of the nodes of zones can be modeled as follows

$$\begin{aligned} P_i(t) =& H_i(t) \cdot (P_{s,0} - 0.820) + 2.84 \times 10^{-4} \cdot \eta \cdot N_{Z,i} \\ &+ 0.549 \cdot \eta \cdot L_{Z,i} \end{aligned} \tag{12}$$

where $i = 1, \cdots, N$, and $H_i(t)$ is the number of active servers in Zone i. $H_i(t)$ is controlled by the data-center level controller.

The CRAC units consume the primary power consumption of the CT system, and the output temperature of the CRAC units is modeled by

$$\dot{T}_{out,i}(t) = -\alpha_i \cdot T_{out,i}(t) + \alpha_i \cdot \min\{T_{in,i}(t), T_{ref,i}(t)\} \tag{13}$$

where $i = N + 1, \cdots, M$, and $T_{ref,i}(t)$ the reference temperature of the CRAC Node i. $T_{ref,i}(t)$ is assumed to be controllable. The min operator in Equation (13) ensures that the supplied air temperatures from CRAC units are not greater than the input temperatures. In the work of Moore et al. (2005), the power consumption of a CRAC node is modeled as follows

$$P_i(t) = \begin{cases} K_i \dfrac{T_{in,i}(t) - T_{out,i}(t)}{COP(T_{out,i}(t))}, & \text{if } T_{in,i}(t) \geq T_{out,i}(t) \\ 0, & \text{if } T_{in,i}(t) < T_{out,i}(t) \end{cases} \tag{14}$$

where $i = N + 1, \cdots, M$, and $COP(T_{out,i}(t))$ the coefficient of performance of CRAC Node i, Moore et al. (2005) shows it is a function of the node's output temperature. Equation (15) demonstrates the relationship between the COP and the output temperatures of the CRAC units. In Equation (8) $\lambda_i = -COP(T_{out,i}(t))$ for CRAC nodes.

$$\begin{aligned} COP(T_{out,i}(t)) =& 0.0068 \cdot T_{out,i}(t)^2 \\ &+ 0.0008 \cdot T_{out,i}(t) + 0.458 \end{aligned} \tag{15}$$

Vectors statement This paper denote the power consumption of thermal nodes of zones with the $M \times 1$ vector $\mathbf{P}_Z(t)$, with $\mathbf{T}_{ref}(t)$ and $\mathbf{P}_C(t)$ denoted the reference temperatures and power consumption of CRAC nodes respectively. We define vector $\mathbf{H}(t)$ as the amount of available servers of zone nodes. The output temperature vector $\mathbf{T}_{out}$ is the state of the thermal dynamics model, $\mathbf{T}_{ref}(t)$ is the controllable input, $\mathbf{P}_Z(t)$ is the uncontrollable input. The input temperature $\mathbf{T}_{in}(t)$ and the power consumption of CRAC units $\mathbf{P}_C(t)$ are outputs of the thermal dynamics model.

The evolution of the thermal dynamics can by presented by

$$\dot{\mathbf{T}}_{out}(t) = A_{T,C}\mathbf{T}_{out}(t) + B_{T,C}\left[\mathbf{P}_Z(t)^T \; \mathbf{T}_{ref}(t)^T\right]^T \quad (16)$$

where $A_{T,C}$ and $B_{T,C}$ can be derived from Equations (11)-(13).

3. CONTROL STRATEGIES

This section considers a hierarchical control strategy, the control architecture is shown in Fig. 2. The data-center level controller collects each data center components' information and provides the optimal set-point to the low level controllers. The low level controllers operate independently from each other in all part of the data center, those works in zones named zone level controller. The controller in each CRAC unit can keep the output air flow temperature up with the reference temperature.

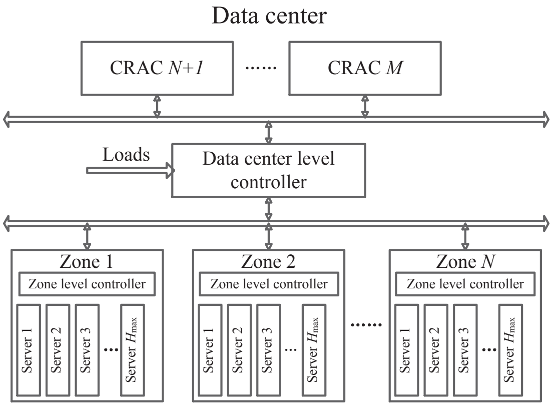

Fig. 2. Hierarchical control of the data center

This paper uses a discrete-time MPC approach to optimize IT and CT resource usage at the data-center level. A load-balancing algorithm is applied in zone level controllers to make the number of connections on the servers as close as possible. At the same time, the zone level controller changes the servers' On-Off states according to the orders of data-center level controller.

3.1 Zone level controller

The zone level controller's main task is to dispatch login requests to available servers in the zone and keep the number of connections on the servers as close as possible. If the upper level controller wants to increase the number of active servers in Zone i, the corresponding zone level controller will choose closed servers and turn them on. If the upper level controller wants to decrease the number of active servers in Zone i, the corresponding zone controller will choose active servers and turn them down. The zone controller will migrate connections on chosen servers to other available servers before closing them.

Quality of service metrics This paper use SNA and SID as metrics of the QoS. Users will receive "Service not available" (SNA) errors if the number of connection servers is insufficient to serve new login requests, and those new login requests will be rejected. The "Service initiated disconnections" (SID) will occurred if a connection server with active users is turned off and users may experience a period of disconnection. When this happens, reconnections will create an artificial surge on the number of new connections, it will generate unnecessary SNAs. Both errors are not allowed to protect user's experience.

Load balancing The load balancing algorithm used in this section was proposed in the work of Chen et al. (2008), where the controller dispatches the following portion of total loads of Zone i to Server δ:

$$p_\delta(t) = \frac{1}{H_i(t)} + \alpha\left(\frac{1}{H_i(t)} - \frac{N_\delta(t)}{N_{Z,i}(t)}\right) \quad (17)$$

where, $\delta = 1, \cdots, H_i(t)$, $i = 1, \cdots, N$, and $\alpha > 0$. α is a parameter that can be tuned to change the dynamic behavior of the system. The algorithm tries to drive the number of connections quickly to uniform for large value of α. This algorithm assigns larger portions to servers with relatively small number of connections. By migrating connections internally the controller can take load off from server δ, then $p_\delta(t)$ can be negative.

Server On-Off state control For the main power consumption of the data center is caused by the IT system, the On-Off state control is used to save more energy. The number of active servers needs to handle the login rate and the total connection of zones. Since a server needs some time to turn on and turn off, we cannot change a server's status whenever we need. The controller recalculates the number of servers that a zone needs for a period of time, and in the time horizon login rates and connections may change a lot. The constraints of active servers of Zone i are considered as:

$$L_{\max} \cdot H_i(t) > \gamma_L \cdot L_{Z,i}(t) \quad (18)$$

$$N_{\max} \cdot H_i(t) > \gamma_N \cdot N_{Z,i}(t) \quad (19)$$

where $\gamma_L > 1$ is a parameter that used to prevent SNAs when $L_{Z,i}(t)$ changes, relatively $\gamma_N > 1$. Both two parameters are associated with the dynamics of $L_{Z,i}(t)$ and $N_{Z,i}(t)$ respectively. The strategies used to dispatch loads in zones also associate with those parameters. Considering the load balancing algorithm used in this paper, when a server in zone j is newly turned on, the login rate signed to it is $(1+\alpha)L_{Z,i}(t)/H_i(t)$, so $\gamma_L > 1+\alpha$.

When a server with active connections on it needs to be turned off, the connections will be migrated at a speed $V_m(t)$ and no new user requests will be dispatched to it. The time-continuous dynamics of the connection on Server δ which need to be turned off is represented as

$$\dot{N_\delta} = \begin{cases} -D_\delta(t) - V_m(t), & \text{if } N_\delta(t) > 0 \\ 0, & \text{if } N_\delta(t) = 0 \end{cases} \quad (20)$$

This server will be turned off immediately when the number of connections on it is zero.

3.2 Data-center level controller

A predictive, discrete-time model of the data center is considered by the data-center level controller. A discrete-time MPC approach is used to optimize the energy efficiency of the total data center with the QoS and the thermal constraints enforced. We consider the optimization problem at a horizon $\Gamma \in \mathbb{N}$, and solve the optimal control problem once in every step. The loads that used to predict the states of the data-center level model are obtained by a short-term load forecasting algorithm.

Load forecasting Let $y(t)$ be the time series under consideration, it can represent $L_{DC}(t)$ or $N_{DC}(t)$ measured at regular time intervals. A auto-regression (AR) model is used to predict the value of $y(t)$ over a period of T time units. The value of $\hat{y}(t)$ measurements as

$$\hat{y}(t) = \sum_{k=1}^{n} a_k \cdot y(t - kT) \tag{21}$$

where n is the order of the AR model, $\{a_k\}$ are parameters of the AR model. In this paper we fix n as 3, and the intervals of prediction is as long as the horizon of optimization problem, which is 6 steps (2 hours).

Optimization problem In the optimization problem we define the predicted value of the variable $\mathbf{N}_Z(t)$ as $\hat{\mathbf{N}}_Z(h|k)$ at the beginning of the hth interval, based on the information available up to the beginning of the kth interval, and similarly we define the variables $\hat{\mathbf{T}}_{in}(h|k)$, $\hat{\mathbf{T}}_{out}(h|k)$, $\hat{\mathbf{T}}_{ref}(h|k)$. The expected value of $q_i(t)$ is denoted with $\hat{\mathbf{q}}(h|k)$ during the kth interval, based on the information available up to kth interval. Similarly, With $\hat{\mathbf{P}}_N(h|k)$ denotes the expected average power consumption of the zones during the kth interval. With $\hat{\mathbf{P}}_C(h|k)$ denotes the expected average power consumption of the CRAC units during the kth interval. With $\hat{\mathbf{H}}(h|k)$ denotes the expected active servers in zones during the kth interval. We define the sets

$$\begin{aligned}
\mathcal{H} &= \left\{\hat{\mathbf{H}}(k|k), \cdots, \hat{\mathbf{H}}(k+\Gamma-1|k)\right\} \\
\mathcal{Q} &= \{\hat{\mathbf{q}}(k|k), \cdots, \hat{\mathbf{q}}(k+\Gamma-1|k)\} \\
\mathcal{T}_{ref} &= \left\{\hat{\mathbf{T}}_{ref}(k|k), \cdots, \hat{\mathbf{T}}_{ref}(k+\Gamma-1|k)\right\} \\
\mathcal{N} &= \left\{\hat{\mathbf{N}}_Z(k|k), \cdots, \hat{\mathbf{N}}_Z(k+\Gamma|k)\right\} \\
\mathcal{L} &= \left\{\hat{\mathbf{L}}_Z(k|k), \cdots, \hat{\mathbf{L}}_Z(k+\Gamma-1|k)\right\}
\end{aligned} \tag{22}$$

Two different control strategies are applied to solve the optimal control problem. In the first one, the controller, which have two independent solver, considers a discrete-time model of the data center and manages the IT and CT in two steps, named uncoordinated MPC. In the first step, the controller solve the optimization problem at time k as below:

$$\begin{aligned}
&\min_{\mathcal{H},\mathcal{Q},\mathcal{N},\mathcal{L}} \sum_{h=k}^{k+\Gamma-1} \left\|\hat{\mathbf{P}}_N(h|k)\right\|_1 \\
&\quad \text{for all } h = k, \cdots, k+\Gamma-1 \\
&\quad \text{computational dynamics} \\
&\text{s.t.} \\
&\quad 0 \le \hat{\mathbf{H}}(h|k) \le \mathbf{H}_{\max} \\
&\quad \mathbf{1}^{\mathrm{T}}\hat{\mathbf{q}}(h|k) = 1,\ \mathbf{0} \le \hat{\mathbf{q}}(h|k) \le \mathbf{1} \\
&\quad \hat{\mathbf{L}}_Z(h|k) = \operatorname{diag}\left\{\mathbf{1}\hat{L}_{DC}(h|k)\right\}\hat{\mathbf{q}}(h|k) \\
&\quad \gamma_L \hat{\mathbf{L}}_Z(h|k) \le L_{\max}\hat{\mathbf{H}}(h|k) \\
&\quad \gamma_N \hat{\mathbf{N}}_Z(h|k) \le N_{\max}\hat{\mathbf{H}}(h|k) \\
&\quad \gamma_N \hat{\mathbf{N}}_Z(h+1|k) \le N_{\max}\hat{\mathbf{H}}(h|k) \\
&\quad \hat{\mathbf{N}}_Z(k|k) = \mathbf{N}_Z(k)
\end{aligned} \tag{23}$$

The controller minimizes the power consumption of the IT system first, and then in the second step, optimize the power consumption of the CT system as follows:

$$\begin{aligned}
&\min_{\mathcal{T}_{ref}} \sum_{h=k}^{k+\Gamma-1} \left\|\hat{\mathbf{P}}_C(h|k)\right\|_1 \\
&\quad \text{for all } h = k, \cdots, k+\Gamma-1 \\
&\quad \text{thermal dynamics} \\
&\text{s.t.} \\
&\quad \mathbf{T}_{ref,min} \le \hat{\mathbf{T}}_{ref}(h|k) \le \mathbf{T}_{ref,\max} \\
&\quad \hat{\mathbf{T}}_{in}(h+1|k) \le \mathbf{T}_{in,\max} \\
&\quad \hat{\mathbf{T}}_{out}(k|k) = \mathbf{T}_{out}(k)
\end{aligned} \tag{24}$$

The second control strategy coordinate IT power consumption and CT power consumption minimization in one optimization problem, named coordinated MPC, it's also based on a discrete-time MPC approach and minimize the power consumption of the data center in each step. At time k, the controller has to solve the following optimization problem:

$$\begin{aligned}
&\min_{\mathcal{H},\mathcal{T}_{ref},\mathcal{Q},\mathcal{N},\mathcal{L}} \sum_{h=k}^{k+\Gamma-1} \left\|\hat{\mathbf{P}}_N(h|k)\right\|_1 + \left\|\hat{\mathbf{P}}_C(h|k)\right\|_1 \\
&\quad \text{for all } h = k, \cdots, k+\Gamma-1 \\
&\quad \text{data center dynamics (5)} - (14) \\
&\text{s.t.} \\
&\quad \mathbf{T}_{ref,min} \le \hat{\mathbf{T}}_{ref}(h|k) \le \mathbf{T}_{ref,\max} \\
&\quad \hat{\mathbf{T}}_{in}(h+1|k) \le \mathbf{T}_{in,\max} \\
&\quad 0 \le \hat{\mathbf{H}}(h|k) \le \mathbf{H}_{\max} \\
&\quad \mathbf{1}^{\mathrm{T}}\hat{\mathbf{q}}(h|k) = 1,\ \mathbf{0} \le \hat{\mathbf{q}}(h|k) \le \mathbf{1} \\
&\quad \hat{\mathbf{L}}_Z(h|k) = \operatorname{diag}\left\{\mathbf{1}\hat{L}_{DC}(h|k)\right\}\hat{\mathbf{q}}(h|k) \\
&\quad \gamma_L \hat{\mathbf{L}}_Z(h|k) \le L_{\max}\hat{\mathbf{H}}(h|k) \\
&\quad \gamma_N \hat{\mathbf{N}}_Z(h|k) \le N_{\max}\hat{\mathbf{H}}(h|k) \\
&\quad \gamma_N \hat{\mathbf{N}}_Z(h+1|k) \le N_{\max}\hat{\mathbf{H}}(h|k) \\
&\quad \hat{\mathbf{T}}_{out}(k|k) = \mathbf{T}_{out}(k),\ \hat{\mathbf{N}}_Z(k|k) = \mathbf{N}_Z(k)
\end{aligned} \tag{25}$$

Because of the first one considers the optimization problem in two steps, the computational complexity of the uncoordinated MPC is lower than the second one. The both control strategies are able to guarantee the enforcement of the QoS and the thermal constraints. The optimal solutions which was generated by the optimization solver, such as $\mathbf{T}_{ref}(k)$, $\mathbf{q}(k)$ and $\mathbf{H}(k)$, will be sent to the low level controllers step by step.

4. SIMULATION RESULT

This paper considers a data center layout like Fig. 1. Server nodes $N1 - N4$ represent a collection of 2 racks, each composed of 30 servers ($H_{\max} = 60$). Servers in $N1$ and $N2$ are slightly more efficient than those in $N3$ and $N4$. The power consumption of idle servers in $N1 - N2$ is 150kW, in $N3 - N4$ is 180kW. The coefficient Φ of the thermal dynamics is

$$\Phi = \begin{pmatrix} 0.01 & 0.5 & 0 & 0.2 & 0.15 & 0.14 \\ 0 & 0.03 & 0 & 0.01 & 0.56 & 0.40 \\ 0 & 0.03 & 0.02 & 0.05 & 0.44 & 0.46 \\ 0 & 0.02 & 0.02 & 0.01 & 0.40 & 0.55 \\ 0.227 & 0.124 & 0.278 & 0.251 & 0.12 & 0.0 \\ 0.335 & 0.114 & 0.268 & 0.163 & 0 & 0.12 \end{pmatrix} \tag{26}$$

Somewhere of the data center doesn't cool efficiently. Because about half of the output hot air of server node $N1$ recycled to server node $N2$, $\phi_{1,2} = 0.5$ is much larger than the amount around it. It means that when server node $N1$ has heavy loads, the input temperature of node $N2$ would be higher than other server nodes. The simulations set the maximum allowed input temperature of every server node at $27°C$, according to ASHRAE Publication (2008): Enviromental guidelines for datacom equipment, expanding the recommended environmental Envelope.

All connection servers are equally in computational performance, the constraints on login rates and active connections are

$$L_{\max} = 70/s, \;\; N_{\max} = 100,000 \tag{27}$$

Before servers being shut down, the corresponding zone level controller migrates the connections on them at a speed $V_m = 100/s$.

Three days of load pattern illustrated in Fig. 3 is taken into consideration. The login rates subject to a random noise uniformly distributed having zero mean and a variance proportional to the mean login rate. The total number of connections fluctuates over day and night.

Simulations were developed under three different control scenarios. In the first scenario the servers in each zone always active and the controller coordinated the energy efficiency of servers and the thermal dynamics in one optimization problem, we named it as constant MPC. Constant MPC is a simplify edition of coordinated MPC. It considered an optimization problem as the same as coordinated MPC, but kept all servers active and ignored On-Off state control. The controller dynamically dispatched jobs to each zone and adjusted the reference temperature for all CRAC units through the simulation. The second scenario used uncoordinated MPC to adjust loads dispatching and the reference temperature of CRAC units dynamically. The third scenario took advantage of coordinated MPC method by considering all the dynamics of the data center. Under all these three scenarios the SID and SNA errors were prevented to influence the user experience.

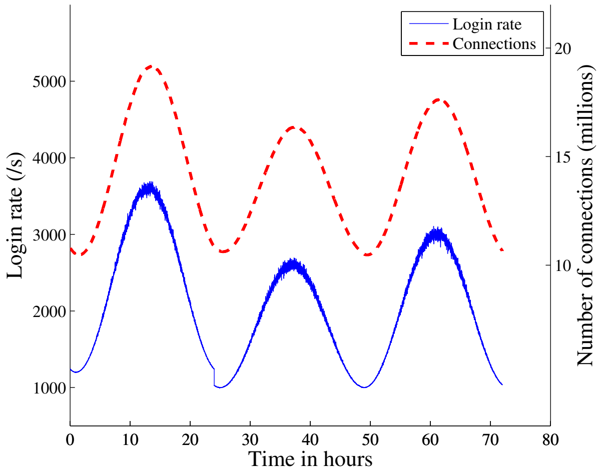

Fig. 3. 3 days of load pattern

The MATLAB Optimization toolbox [Coleman et al. (1999)] was used as the numerical solver. We assume that all zone nodes have the same initial state. The time step of the simulation is 30 s and the MPC controller solves the optimization problem every 20 min. The prediction horizon of the optimization problem is 6 steps (2 hour), and the control horizon is 3 steps (1 hour). The servers were forced to process all login requests, and no active connections were dropped due to the migration actions. The energy consumption of the total data center under different control strategies are listed in Table 1. Compared with the constant MPC, the coordinated MPC method reduced the energy consumption of the total data center to 27.58% and the uncoordinated MPC to 24.52% with enforced the QoS and the thermal constraints when considering dynamic server provisioning. There are no SNA and SID errors occurred in the simulation processes. The energy saving under coordinated MPC was main come from the CT system compared with it under uncoordinated MPC. Benefit from took thermal dynamics into consideration in the optimization of active server numbers, the PUE decreased to 1.24 under coordinated MPC. Uncoordinated MPC allocated login requests to zones regardless of thermal dynamics and have a relatively high PUE almost to 1.34.

Table 1. Comparing of Energy Consumption (kWh) Under Different Control Scenarios

Control Scenarios	IT Energy	CT Energy	Total Energy	PUE
Constant MPC	3066.20	993.98	4060.18	1.32
Uncoordinated MPC	2293.34	771.43	3064.77	1.34
Coordinated MPC	2380.02	560.23	2940.25	1.24

The power consumption of total data center under different control scenarios is shown in Fig. 4.

Fig. 5 indicates that the variation of the number of active servers in nodes $N1 - N4$ under MPC, the number of active servers reflects the amount of login requests each zone processes. The uncoordinated MPC strategy dispatched loads to server Node $N1 - N2$ first and $N3 - N4$ last, because the servers in Node $N1$ and $N2$ have relatively high energy efficiency.

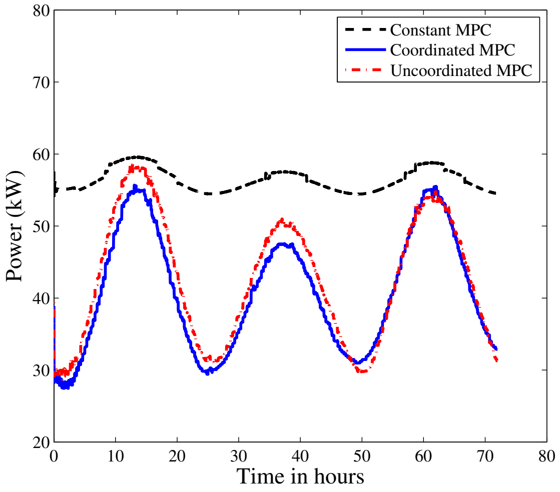

Fig. 4. Total data center power consumption

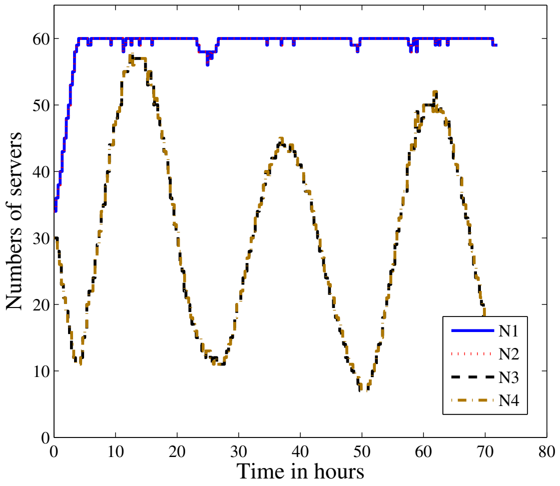

Fig. 5. The number active servers in zone $N1 - N4$ under Uncoordinated MPC

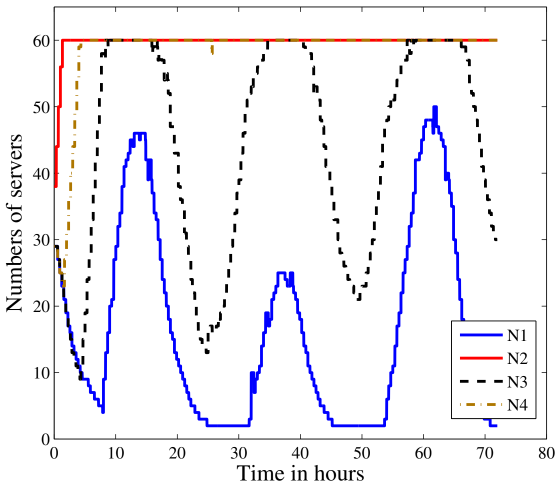

Fig. 6. The number active servers in zone $N1 - N4$ under Coordinated MPC

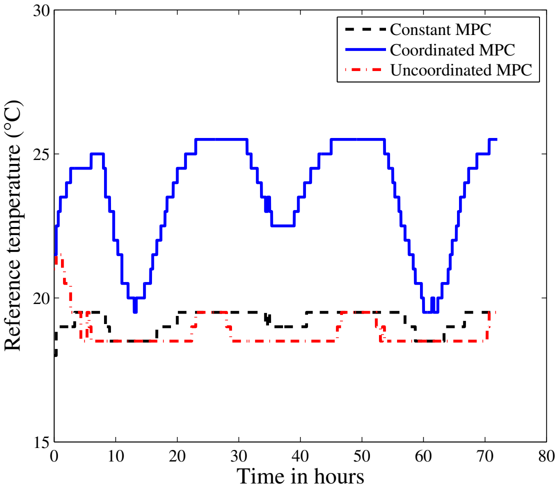

Fig. 7. Average reference temperature

Fig. 6 indicates data-center level controller using coordinated MPC dispatched loads to server Node $N2$ first and $N1$ last, where the servers in Node $N2$ have relatively high energy efficiency and the output air flow recirculated less. Because about half of the output hot air from server Node $N1$ recirculated to server Node $N2$, $N1$ have the last priority to process loads.

Fig. 7 demonstrates the average CRAC reference temperature under three control strategies. The average reference temperature varies a lot under coordinated MPC and always higher than uncoordinated MPC case. When compared with uncoordinated MPC, coordinated MPC is able to maintain a higher level of cooling efficiency and is able to largely reduce the power consumption of the CRAC units. Although uncoordinated MPC largely reduce the energy consumption of IT system, but the average reference temperature under uncoordinated MPC is lower than which under constant MPC. It's because uncoordinated MPC used N1 as one of the main zones to process jobs. The recirculation of the output air flow of Node N1 caused the CRAC units turned down reference temperatures to enforce the thermal constraints.

Fig. 8 shows coordinated MPC and uncoordinated MPC get energy saved from both IT system and CT system. The power consumption of IT and CT have the same variation, and them changed more quickly compared with constant MPC scenario. In simulations, three different control scenarios were able to enforce the inlet temperature constraints and prevented SNA and SID errors.

In the simulation processes, because of the computational complexity, the solver consumes more time to solve the optimization problem (25) than the sum of optimization problem (23)(24) in our simulation environment, and both of coordinated MPC and uncoordinated MPC satisfied the requirement of real-time control, and after solving the optimization problem there was enough time to change servers' states. We considered another case where the servers in $N3$ and $N4$ are slightly more efficient than those in $N1$ and $N2$. The power consumption of idle servers in $N3 - N4$ is 150kW, in $N1 - N2$ is 180kW. Then no matter

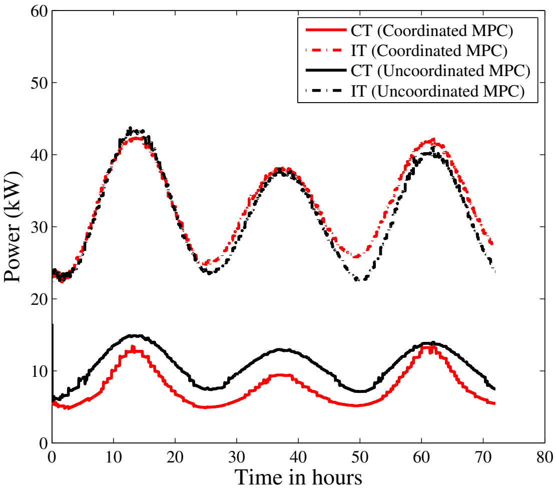

Fig. 8. Power consumption of the IT system and the CT system

under coordinated MPC or uncoordinated MPC the N1 always has the last priority to process login requests. The cooling of the data center would be much better than before. The energy consumption of the total data center under different control strategies are listed in Table 2. Both Coordinated MPC and uncoordinated MPC got a significant energy saving compared with constant MPC. Considering the efficiency of the control algorithm, the uncoordinated MPC is preferred in this case.

Table 2. Comparing of Energy Consumption (kWh) Under Different Control Scenarios

Control Scenarios	IT Energy	CT Energy	Total Energy	PUE
Constant MPC	3066.20	1121.09	4187.29	1.36
Uncoordinated MPC	2208.65	599.12	2807.76	1.27
Coordinated MPC	2252.93	557.53	2810.46	1.24

Coordinated MPC could achieve more energy saving than uncoordinated MPC when the data center's relatively high efficient server zones have large hot air recirculation. But for a data center with relatively high server zone cooling efficiency ,the uncoordinated MPC is preferred. By taking dynamic server provisioning into optimal control, both control strategies could obtain more than 24% energy reduction.

5. CONCLUSIONS

A model including zone level dynamics and data-center level dynamics is introduced in this paper, and three control strategies are used in the simulation section, constant MPC, uncoordinated control and coordinated control. Three aspects, dynamic server provisioning, server efficiency and thermal dynamics, are considered in the optimization problem. By dynamically changing active server numbers, reference temperatures and dispatching loads to different server zones, the energy efficiency of the data center was improved significantly. The simulation result indicated that coordinating dynamic server provisioning and thermal dynamics in an optimization process is possible to reduce the energy consumption of the total data center. The comparison of three cases provides an example to achieve energy saving in data centers which have cooling deficiency.

In the simulation process, in order to keep the quality of service, the accurateness of the forecast of data center loads is very important for the MPC controller. Dynamic server provisioning depending on various types of loads and services is necessary to be studied in the future. The primary goal of the control strategy is to minimize the energy consumption of the total data center. More other factors that are important in operating a data center could be added to the cost function.

REFERENCES

ASHRAE Publication (2008). Enviromental guidelines for datacom equipment, expanding the recommended environmental envelope. Technical report, American Society of Heating, Refrigerating and Air-Conditioning Engineering.

Chase, J.S., Anderson, D.C., Thakar, P.N., Vahdat, A.M., and Doyle, R.P. (2001). Managing energy and server resources in hosting centers. *ACM SIGOPS Operating Systems Review*, 35(5), 103–116.

Chen, G., He, W., Liu, J., Nath, S., Rigas, L., Xiao, L., and Zhao, F. (2008). Energy-aware server provisioning and load dispatching for connection-intensive internet services. In *USENIX Symposium on Networked Systems Design and Implementation*, volume 8, 337–350.

Coleman, T., Branch, M.A., and Grace, A. (1999). *Optimization toolbox for use with MATLAB: user's guide, version 2.* Math Works, Incorporated.

Doyle, R.P., Chase, J.S., Asad, O.M., Jin, W., and Vahdat, A. (2003). Model-based resource provisioning in a web service utility. In *USENIX Symposium on Internet Technologies and Systems*, volume 4, 5–5.

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*, 61–75.

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(1), 254–268.

Parolini, L., Sinopoli, B., and Krogh, B. (2008). Reducing data center energy consumption via coordinated cooling and load management. In *Proceedings of the 2008 Conference on Power Aware Computing and Systems, HotPower*, volume 8, 14–14.

Parolini, L., Sinopoli, B., and Krogh, B. (2011). Model predictive control of data centers in the smart grid scenario. In *Proceedings of the 18th International Federation of Automatic Control (IFAC) World Congress*, volume 18, 10505–10510.

Tang, Q. (2008). *Thermal-aware scheduling in environmentally coupled cyber-physical distributed systems.* Ph.D. thesis, Arizona State University.

Tang, Q., Mukherjee, T., Gupta, S.K.S., and Cayton, P. (2006). Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters. In *The Fourth International Conference on Intelligent Sensing and Information Processing (ICISIP)*, 203–208.