A Cyber-Physical Systems Approach to Data Center Modeling and Control for Energy Efficiency

Luca Parolini, *Student Member, IEEE,* Bruno Sinopoli, *Member, IEEE,* Bruce H. Krogh, *Fellow, IEEE,* and Zhikui Wang, *Member, IEEE*

***Abstract*—This paper presents data centers from a cyber-physical system (CPS) perspective. Current methods for controlling information technology (IT) and cooling technology (CT) in data centers are classified according to the degree to which they take into account both cyber and physical considerations. To evaluate the potential impact of coordinated CPS strategies at the data-center level, we introduce a control-oriented model that represents the data center as two coupled networks: a *computational network* representing the cyber dynamics and a *thermal network* representing the physical dynamics. These networks are coupled through the influence of the IT on both networks: servers affect both the quality of service (QoS) delivered by the computational network and the generation of heat in the thermal network. Using this model, three control strategies are evaluated with respect to their energy efficiency and computational performance: a *baseline strategy* that ignores CPS considerations, an *uncoordinated strategy* that manages the IT and CT independently, and a *coordinated strategy* that manages the IT and CT together to achieve optimal performance with respect to both QoS and energy efficiency. Simulation results show that the benefits to be realized from coordinating the control of IT and CT depend on the distribution and heterogeneity of the computational and cooling resources throughout the data center. A new *cyber-physical index* (CPI) is introduced as a measure of this combined distribution of cyber and physical effects in a given data center. We illustrate how the CPI indicates the potential impact of using coordinated CPS control strategies.**

***Index Terms*—Data centers, Cyber-Physical systems, Energy Efficiency, Thermal management, Computer applications, Predictive control.**

I. Introduction

Data centers are facilities hosting a large number of servers dedicated to massive computation and storage. They can be used for several purposes, including interactive computation (e.g., web browsing), batch computation (e.g., renderings of images and sequences), or real-time transactions (e.g., banking). Data centers can be seen as a composition of information technology (IT) systems and a support infrastructure. The IT systems provide services to the end users while the infrastructure supports the IT systems by supplying power and cooling. IT systems include servers, storage and networking devices, middleware and software stacks, such as hypervisors, operating systems, and applications. The support infrastructure includes backup power generators, uninterruptible power supplies (UPSs), power distribution units (PDUs), batteries, and power supply units that generate and/or distribute power to the individual IT systems. The cooling technology (CT) systems, including server fans, computer room air conditioners (CRACs), chillers, and cooling towers, generate and deliver the cooling capacity to the IT systems [1]–[6].

To provide the quality of service (QoS) required by service level agreements, the IT control system can dynamically provision IT resources or actively manage the workloads through mechanisms such as admission control and workload balance. The IT systems consume power and generate heat whenever they are on. The power demand of the IT system can vary over time and is satisfied by the power delivery systems. The CT systems extract the heat to maintain the thermal requirements of the IT devices in terms of temperature and humidity. The IT, power and cooling control systems have to work together to manage the IT resources, power and cooling supply and demand.

The number of data centers is rapidly growing throughout the world, fueled by the increasing demand for remote storage and cloud computing services. Fig. 1 shows the increase in data center expenditures for power, cooling, and new servers from 1994 until 2010. Energy consumed for computation and cooling is dominating data center run-time costs [7], [8]. A report of the U.S. Environmental Protection Agency (EPA) shows that data center power consumption doubled from 2000 to 2006, reaching a value of 60 TWh/yr (terawatt hour/year) [9]. Historical trends suggest another doubling by the end of 2011.

As computational density has been increasing at multiple levels, from transistors on integrated circuits (ICs), to servers in racks, to racks in a room, the rate at which heat must be removed has increased, leading to nearly equal costs for operating the IT system and the CT system [10], [11]. Fig. 2 shows the measured and the expected growth of the power consumption density in data center equipment from 1994 until 2014 [12]. Available cooling capacity has in some cases become the limiting factor on the computational capacity [13]. Although liquid cooling is a promising alternative to air cooling, particularly for high-density data centers, this technology has not been widely adopted so far due to high costs and safety concerns [14].

The increase in data center operating costs is driving innovation to improve their energy efficiency. A measure of data center efficiency typically used in industry is the *power usage*

L. Parolini, B. Sinopoli, and B. H. Krogh are with the Department of Electrical and Computer Engineering of Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA (e-mail: {lparolin|brunos|krogh}@ece.cmu.edu).

Z. Wang is with the Sustainable Ecosystem Research Group of HP Labs, 1501 Page mill road ms 1183, Palo Alto, CA 94304 (e-mail: zhikui.wang@hp.com)

This work was supported in part by the National Science Foundation, Grant No. ECCS 0925964.

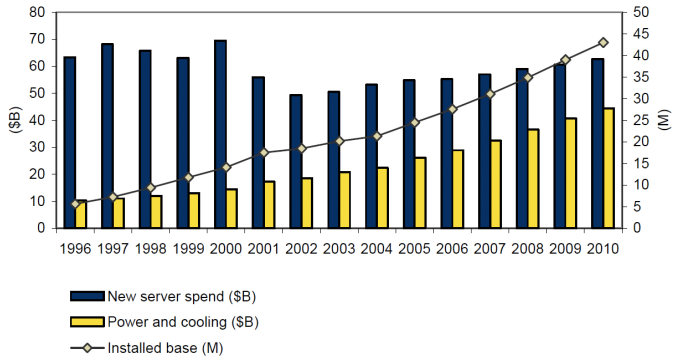


Fig. 1. Data center spending trend [15].

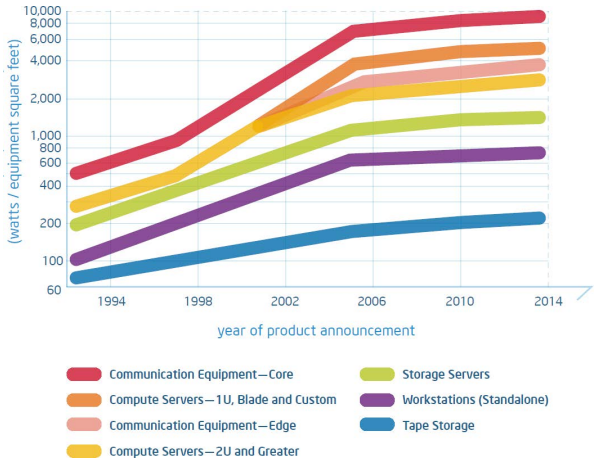


Fig. 2. Power consumption per equipment footprint [12].

effectiveness (PUE), defined as the ratio between the total facility power consumption over the power consumption of the IT [16]. A PUE of 1.0 would indicate that all of the data center power consumption is due to the IT. Fig. 3 shows the PUE values measured by 60 different data centers in 2007 [16]. Their average PUE value is 2.03, i.e., almost half of the total data center power consumption is consumed by the CT, which dominates the non-IT facility power consumption. Data centers based on state-of-the-art cooling and load balancing technology can reach PUE values of 1.1, i.e., 90.9% of the total data center power consumption is consumed by IT.[1] A drawback of PUE is that it does not take into account IT equipment efficiency.

This paper considers data centers as cyber-physical systems (CPS), with a focus on run-time management and operating costs. The following section reviews current methods for controlling information technology (IT) and cooling technology (CT) in data centers, noting to the degree to which they take into account both cyber and physical considerations. To evaluate the potential impact of coordinated CPS strategies at the data-center level, Sec. III introduces a control-oriented model that represents the data center as two coupled networks: a *computational network* representing the cyber dynamics and a *thermal network* representing the physical dynamics. These

[1]http://www.google.com/corporate/datacenter/efficiency-measurements.html

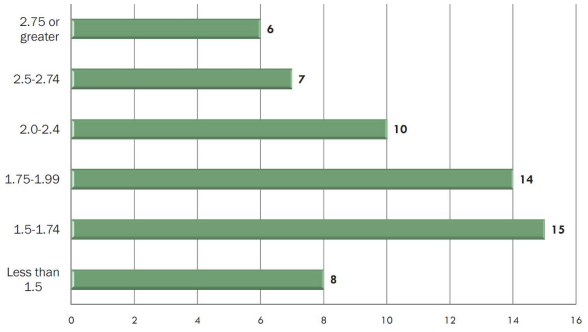


Fig. 3. PUE measurements and number of respondents. Average value is 2.03 [16].

networks are coupled through the influence of the IT on both networks: servers affect both the quality of service (QoS) delivered by the computational network and the generation of heat in the thermal network. Three representative data center level control strategies are defined in Sec. IV: a *baseline strategy* that ignores CPS considerations, an *uncoordinated strategy* that manages the IT and CT independently, and a *coordinated strategy* that manages the IT and CT together to achieve optimal performance with respect to both QoS and energy efficiency. Simulation results presented in Sec. V show that the benefits to be realized from coordinating the control of IT and CT depend on the distribution and heterogeneity of the computational and cooling resources throughout the data center. As a measure of this combined distribution of cyber and physical effects in a given data center, a new *cyber-physical index* (CPI) is introduced in Sec. VI and we illustrate how the CPI indicates the potential impact of using coordinated CPS control strategies. The paper concludes with a summary of the insights gained by viewing data centers from a CPS perspective and the applicability of these observations to other cyber-physical systems.

II. Trends in Data Center Control

This section reviews previous work on control in data centers. Active IT management, power control and cooling control systems can each have their own hierarchies [17], [18], and the interactions and coordination between the cyber and physical aspects of data centers occur in multiple spatial and temporal scales as well. Depending on both the metrics and the control variables that are used, we classify the work into three typical scales: *server level*, *group level*, and *data center level*.

At every level, controllers can base their control actions considering only the state of the local controlled subsystem, or they can take into consideration the effects on other subsystems. In the first case, the controller acts based on a local-view of the system, whereas in the second case, the controller acts based on a global-view of the system. We classify the control approaches based on the scale of the controlled subsystem. For example, a controller at the server level manages the operations of a single server, even though the way the control actions are taken may derive from a global-view of the data center.

A. Server level control

There are many control variables available at the server level for IT, power and cooling management. The "server" in this case means all the IT devices, including the computing servers, the storage units and the networking equipment. The computing resources, such as CPU cycles, memory capacity, storage access bandwidth and networking bandwidth, are all local resources that can be dynamically tuned, especially in a virtualized environment. Power control can be done from either the demand side or the supply side, even at the server level. The power consumption of servers can be controlled by active management of the workload hosted by the server, for instance, through admission control, load balance, and by workload migration/consolidation. On the other hand, the power consumption can be tuned through physical control variables such as dynamic voltage and frequency scaling (DVFS) and through the ON-OFF state control [19]–[27]. DVFS has been implemented already in many operating systems, e.g., the "CPU governors" in Linux systems. CPU utilization usually drives the DVFS controller, which adapts the power consumption to the varying workload.

Previous work has been focused on how to deal with the tradeoff between the power consumption and IT performance. For instance, Varma *et al.* [28] discuss a control-theoretic approach to DVFS. Cho *et al.* [19] discuss a control algorithm that varies both the clock frequency of a microprocessor and the clock frequency of the memory. Leverich *et al.* [29] propose a control approach to reduce static power consumption of the chips in a server through dynamic per-core power gating control.

Cooling control at the server level is usually implemented through active server fan tuning to cool down the servers [30]. Similar to power management, the thermal status of the servers, e.g., the temperatures of the processors, can be affected by active control of the workload or power as well. As one example, Mutapcic *et al.* [24] focus on the maximization of the the processing capabilities of a multi-core processor subject to a given set of thermal constraints. In another example, Cohen *et al.* [23] propose control strategies to control the power consumption of a processor via DVFS so as to enforce the given constraints on the chip temperature and on the workload execution.

B. Group level control

There are several reasons to control groups of servers rather than a single servers. First, the IT workload nowadays most often runs on multiple nodes, e.g., a multi-tier application can span a set of servers. Second, when the metrics that drive the controllers are the performance metrics of the IT workloads, the control decision has to be made at the group level. One typical example is the workload migration/consolidation in the virtualized environment, which is applied widely these days for sharing the resources, improving resource utilization and reducing power consumption [31]–[37]. Third, the IT and the infrastructure are usually organized into groups, e.g., the server enclosures that contain several servers cooled by a set of fans, the racks that have more than forty servers each, the rows of racks, and the hot/cold aisles. In any case, the principal goal of group-level control is still to meet the cyber performance requirements while improving the physical resource utilization and energy efficiency. A few examples follow.

Padala *et al.* [38] propose a control algorithm to allocate IT resources among servers when these are divided among different tiers. Gandhi *et al.* [25], [26] focus on workload scheduling policies and transitions between the OFF and ON states of servers, as a means to minimize the average response time of a data center given a certain power budget. A model predictive control (MPC) [39], [40] approach is considered in the work of Aghajani *et al.* [27], where the goal of the control action is to dynamically adjust the number of active servers based on the current and predicted workload arrival rates. Wang *et al.* [34] propose a model predictive controller to minimize the total power consumption of the servers in an enclosure subject to a given set of QoS constraints. Tolia *et al.* [33] discuss an MPC approach for coordinated workload performance, server power and thermal management of a server enclosure, where a set of blade servers shares the cooling capacity of a set of fans. The fans are controlled through a multi-input multi-output (MIMO) controller to minimize the aggregate fan power, while the location-dependent cooling efficiency of the fans is exploited so that the decisions of workload migration can result in least total server and fan power consumption.

C. Data center level control

Depending on the boundaries defining the groups, group-level controllers have to be implemented at the data center scale in some cases. For instance, in a workload management system, the workload may be migrated to the servers at any location in a data center. The other reason for the need of data center level control is that power and cooling capacity has to be shared throughout the data center. As one example, the cooling capacity in a raised-floor air-cooled data center can be generated by a chiller plant, distributed to the IT systems through a set of CRAC units, the common plenum under the floor, and the open space above the floor. Sharing the capacity makes the cooling management in the first order a data center level control [41]–[43].

Quershi *et al.* [44] discuss the savings that can be achieved by migrating workload to the data centers locations where the electricity cost is at its minimum. A similar problem is considered in the work of Rao *et al.* [45]. Tang *et al.* [46] discuss a control algorithm that allocates the workload among servers so as to minimize their peak inlet temperatures. Parolini *et al.* [47], [48] consider a data center as a node of the smart-grid, where time-varying and power-consumption-dependent electricity price information is used to manage data center operations.

Bash *et al.* [41] discuss a control algorithm for the CRAC units. The proposed control approach aims at minimizing the amount of heat removed by each CRAC unit, while enforcing the thermal constraints of the IT. Anderson *et al.* [49] consider a MIMO robust control approach to the control of CT. Computational fluid dynamic (CFD) simulations are a widely used tool to simulate and predict the heat distribution in a

data center. Such simulations take a very long time to execute and cannot be used in real-time control applications. In [50], Toulouse *et al.* discuss an innovative approach to CFD which is able to perform fast simulations.

Chen *et al.* [51] propose a holistic workload, power and cooling management framework in virtualized data centers through exploration of the location-dependent cooling efficiency in the data center level. Workload migration and consolidation through virtual machine (VM) migration is taken as the main approach for active power management. On the thermal side, the rack inlet temperatures are under active control of a cooling controller. The controller dynamically tunes the cooling air flow rates and temperatures from the CRAC units, in response to the "hot-spots" of the data center. On the cyber side, the authors consider multi-tier IT applications or workloads hosted in VMs that can span multiple servers, and be able to migrate VMs. The models are developed for application performance such as end-to-end response time. Based on the models and the predicted workload intensity, the computing resource demand of the workloads can be estimated. The demand of the VMs hosted by a server is then satisfied through dynamic resource allocation, if possible. Otherwise, the VMs may be migrated to other servers that have resources available. When possible, the VMs are consolidated onto fewer servers so that the idle ones can be turned off to save power. With the decision to migrate workload and turn on/off servers by the consolidation controller, the effect of the actions on the cooling power is taken into consideration through introduction of the *local workload placement index* (LWPI). As one index for the interaction between the cyber and the physical systems, the LWPI indicates how much cooling capacity is available in a certain location of the data center, and how efficiently a server in the location can be cooled. Using the LWPI, the consolidation controller tends to migrate the workload into the servers that are more efficiently cooled than others, while turning off the idle servers that are located at "hotter spots".

An idle server typically consumes about 60% of its peak power. Servers in data centers typically operate between 10% and 50% of their maximum utilization and often are completely idle [7], [52], [52]–[54]. In order to maximize the server energy efficiency it would be desirable to operate them at 100% utilization by consolidating the workload onto a few servers and turning off the unused ones. The main problem related to turning off a server is the time required to turn it back on. This time, often called *setup time*, is on the order of a few minutes and it is typically is not acceptable for interactive workload [26], [27]. This problem can be solved by predicting the incoming workload and adjusting the computational power accordingly. Another potential drawback of using such techniques is the effect they may have on the cooling effort. Concentrating computation on a few servers while turning off others has in general the effect of generating hot spots for which the cooling system needs to do extra work. As air cooling cannot be targeted to a specific device this method may in turn over-cool the overall data center, leading to energy inefficiencies in the cooling system, thus potentially offsetting the savings achieved by reducing overall server power consumption.

III. A CPS MODEL FOR DATA CENTERS

As described in the previous section, the CT is managed at the data center level. At this level, the thermal properties of the IT are managed as groups of components (racks of servers) aggregated into *zones*. From a CPS perspective, the CT is purely physical, consuming power to removing heat from the zones, whereas zones have both cyber and physical characteristics. The IT processes the computational workload and also consumes power and generates heat. In this section, we present a model of the data center level dynamics using common representations of these cyber and physical features as well as the coupling between them. We use this model in the following sections to study the aspects of data center architectures and operating conditions that influence the potential impact of coordinated CPS (IT and CT) control strategies.

We model the data center as two networks. In the *computational network*, nodes correspond to the cyber aspects of zones, which can process jobs at rates determined by the allocated resources, and connections represent pathways for moving jobs between zones. The *thermal network* includes nodes to represent the thermal aspects of zones along with nodes for the CT. Connections in the thermal network represent the exchange of heat between nodes as determined by the data center's physical configuration. The two networks are coupled by connections between the two nodes that represent each zone in the computational network and thermal network, respectively. These connections reflect the correlation between the computational performance of the resources allocated in the zone (a control variable) and the physical features of the zone (power consumption and generated heat). The following subsections elaborate on the details of the dynamics for each network.

A. Computational network

We model the computational network using a fluid approximation of the workload execution and arrival processes, i.e., the workload is represented by job flow rates rather than as discrete jobs. The proposed modeling approach represents a first-order approximation of a queuing system. The strength of the proposed approach resides in its simplicity, which allows an uncluttered discussion of the model and of the control approach. On the other hand, a certain amount of computational details of a data center are not included in the model. We consider the approximation provided by the proposed model adequate for the goal of the paper. However, when additional details of the system are relevant for the system analysis, more refined approximations can be considered [55].

Let N be the number of nodes and let $l_i(\tau)$ denote the amount of workload in the i^{th} node at time τ. From a queuing perspective, $l_i(\tau)$ represents the queue length of the i^{th} node at time τ. The workload arrival rate at the data center is denoted by $\lambda^W(\tau)$. The relative amount of workload that is assigned to the i^{th} node at time τ is denoted by $s_i(\tau)$. Variables $\{s_i(\tau)\}$ represent the workload scheduling (or allocation) action. The rate at which workload migrates from the i^{th} computational

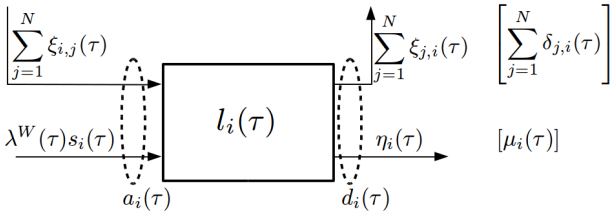


Fig. 4. Input, state, and output variables of the i^{th} computational node.

node to the j^{th} computational node at time τ is denoted by $\xi_{j,i}(\tau)$. The rate at which the workload departs from the i^{th} node at time τ, after being executed on the node, is denoted by $\eta_i(\tau)$. Let $\mu_i(\tau)$ denote the desired workload execution rate at the i^{th} node at time τ and let $\delta_{j,i}(\tau)$ denote the required migration rate of workload from the i^{th} computational node to the j^{th} computational node at time τ. Variables $\{\mu_i(\tau)\}$, $\{s_i(\tau)\}$, and $\{\delta_{j,i}(\tau)\}$ are controllable variables, whereas $\{l_i(\tau)\}$ are the state variables. Fig. 4 illustrates these variables that describe the i^{th} computational node (a zone in the data center).

We define the following variables for $i = 1, \ldots, N$:

$$\begin{aligned} a_i(\tau) &= \lambda^W(\tau)s_i(\tau) + \sum_{j=1}^{N} \xi_{i,j}(\tau), \\ d_i(\tau) &= \eta_i(\tau) + \sum_{j=1}^{N} \xi_{j,i}(\tau), \\ \nu_i(\tau) &= \mu_i(\tau) + \sum_{j=1}^{N} \delta_{j,i}(\tau). \end{aligned}$$

Variable $a_i(\tau)$ represents the total rate at which workload arrives at the i^{th} node, variable $d_i(\tau)$ represents the total rate at which workload departs from the i^{th} node, and variable $\nu_i(\tau)$ represents the desired total workload departure rate from the i^{th} node. The evolution of the amount of workload at the i^{th} computational node is given by

$$\dot{l}_i(\tau) = a_i(\tau) - d_i(\tau). \tag{1}$$

The relationship between the departure rate, the state, and the control variables at the i^{th} node is given by

$$\eta_i(\tau) = \begin{cases} \mu_i(\tau) & \text{if } l_i(\tau) > 0 \text{ or } a_i(t) > \mu_i(t) \\ a_i(t) & \text{otherwise} \end{cases}. \tag{2}$$

The relationship between the workload migration rate, the state, and control variables at the i^{th} node can be written as

$$\xi_{j,i}(\tau) = \begin{cases} \delta_{j,i}(\tau) & \text{if } l_i(\tau) > 0 \text{ or } a_i(t) > \nu_i(t) \\ \dfrac{\delta_{j,i}(\tau)}{\sum_{j=1}^{N} \delta_{j,i}(\tau)} (a_i(\tau) - \eta_i(\tau)) & \text{otherwise} \end{cases}. \tag{3}$$

Eqs. (2) and (3) model the case where the i^{th} node does not migrate workload to other computational nodes if the total rate of incoming workload is lower than, or equal to, the desired workload execution rate, i.e., $a_i(\tau) \leq \mu_i(\tau)$ and the queue length is 0, i.e., $l_i(\tau) = 0$.

The model for the workload execution rates developed above is sufficient for our purposes, but we note that it can be extended to include different migration policies, workload classes, hardware requirements, and interactions among different types of workloads [47], [48], [56].

B. Thermal network

Let M be the number of nodes in the thermal network. The dynamics of the thermal network is characterized in terms of temperatures associated with each of the thermal nodes. For each node of the network we define two temperatures: the *input* temperature and the *output* temperature. The input temperature of the i^{th} node represents the amount of heat received from the other thermal nodes and its value at time τ is denoted by $T_{\text{in},i}(\tau)$. Variable $T_{\text{in},i}(\tau)$ includes the recirculation and cooling effects due to all thermal nodes. The output temperature of the i^{th} thermal node, denoted by $T_{\text{out},i}(\tau)$, represents the amount of heat contained in the i^{th} thermal node at time τ.

Following Tang *et al.* [57], we assume each input temperature is a linear combination of the output temperatures from all of the nodes, that is,

$$T_{\text{in},i}(\tau) = \sum_{j=1}^{M} \psi_{i,j} T_{\text{out},j}(\tau), \qquad \text{for all } i = 1, \ldots, M, \tag{4}$$

where the coefficients $\{\psi_{i,j}\}$ are non-negative and $\sum_{j=1}^{M} \psi_{i,j} = 1$ for all $i = 1, \ldots, M$. The values of the coefficients $\{\psi_{i,j}\}$ can be estimated following the procedure in [57]. We collect the input and the output temperatures in the $M \times 1$ vectors $\boldsymbol{T}_{\text{in}}(\tau)$ and $\boldsymbol{T}_{\text{out}}(\tau)$ respectively. Consequently their relationship can be written in vector form as

$$\boldsymbol{T}_{\text{in}}(\tau) = \Psi \boldsymbol{T}_{\text{out}}(\tau), \tag{5}$$

where $\{\psi_{i,j}\}$ are the components of the matrix Ψ.

Measurements taken on a server in our laboratory and discussed in [58] show that a linear time-invariant (LTI) system is a good approximation of the evolution of the outlet temperature of a server. Therefore, we model the evolution of the output temperatures of the thermal nodes for zones by

$$\dot{T}_{\text{out},i}(\tau) = -\mathsf{k}_i T_{\text{out},i}(\tau) + \mathsf{k}_i T_{\text{in},i}(\tau) + \mathsf{c}_i \mathsf{p}_i(\tau), \tag{6}$$

where $\frac{1}{\mathsf{k}_i}$ is the time constant of the temperature of i^{th} node, c_i is the coefficient that maps power consumption into output temperature variation, and $\mathsf{p}_i(\tau)$ is the power consumption of the i^{th} node at time τ.

The power consumption of the nodes representing a zone is proportional to the rate at which workload departs, after being executed, from the associated computational node

$$\mathsf{p}_i(\tau) = \alpha_i \eta_i(\tau), \tag{7}$$

where α_i is a non-negative coefficient. A linear model is chosen since we assume that lower-level controllers, for example, the one proposed by Tolia *et al.* [33], can be used to make the power consumption of a zone proportional to the amount of workload processed by the zone. This linear model can be extended to include more complicated functions which may account for the ON-OFF state of every server.

In the CT we focus on the CRAC units, which are the primary power consumers. The output temperatures of the CRAC units are modeled by

$$\dot{T}_{\text{out},i}(\tau) = -\mathsf{k}_i T_{\text{out},i}(\tau) + \mathsf{k}_i \min\{T_{\text{in},i}(\tau), T_{\text{ref},i}(\tau)\}, \quad (8)$$

where $T_{\text{ref},i}(\tau)$ represents the reference temperature of the CRAC node i and is assumed to be controllable. The $\min$ operator in (8) ensures that the node always provides output temperatures that are not greater than the input temperatures. As discussed by Moore *et al.* [59], the power consumption of a CRAC node is given by

$$\mathsf{p}_i(\tau) = \begin{cases} \mathsf{c}_i \dfrac{T_{\text{in},i}(\tau) - T_{\text{out},i}(\tau)}{COP(T_{\text{out},i}(\tau))} & T_{\text{in},i}(\tau) \geq T_{\text{out},i}(\tau) \\ 0 & T_{\text{in},i}(\tau) < T_{\text{out},i}(\tau), \end{cases} \quad (9)$$

where c_i is a coefficient that depends on the amount of air passing through the CRAC and the air heat capacity. The variable $COP(T_{\text{out},i}(\tau))$ is the *coefficient of performance* of the CRAC unit modeled by the i^{th} node, which is a function of the node's output temperature [59].

In order to provide a compact representation of the overall model we use vector notation. We denote with the $N \times 1$ vector $\mathbf{p}_{\mathcal{N}}(\tau)$ the power consumption of the thermal nodes representing the zones, and with $\mathbf{p}_{\mathcal{C}}(\tau)$ and $\boldsymbol{T}_{\text{ref}}(\tau)$, respectively, the power consumption and the reference temperatures of the thermal nodes representing CRAC units. The state of the thermal network is represented by the vector $\boldsymbol{T}_{\text{out}}(\tau)$, the controllable input of the thermal network by the vector $\boldsymbol{T}_{\text{ref}}(\tau)$, and the uncontrollable input of the thermal network by $\mathbf{p}_{\mathcal{N}}(\tau)$. Finally, the outputs of the thermal network are the vector of the thermal node input temperatures $\boldsymbol{T}_{\text{in}}(\tau)$ and the vector $\mathbf{p}_{\mathcal{C}}(\tau)$. The vector $\boldsymbol{T}_{\text{in}}(\tau)$ is a function of the network state and therefore, it is an output of the network. However, when we look at a single node, the input temperature becomes an uncontrollable input of the node. In this sense, the input vector is an output of the thermal network and, at the same time, an uncontrollable input for each of the nodes.

IV. Data Center Control Strategies

Three control strategies are introduced in this section: *baseline*, *uncoordinated*, and *coordinated*. The control strategies are abstractions of three different control approaches that can be implemented at the data center level. The baseline strategy represents those control approaches where IT and CT are set so as to satisfy the QoS and the thermal constraints for the worst-case scenario, regardless of the actual computational and cooling demands of the data center. The uncoordinated strategy represents those control approaches where the efficiency of IT and CT are considered in two separate optimization problems. The coordinated strategy represents those control approaches where the efficiencies IT and CT are controlled using a single optimization problem.

The goal of the control strategies is to minimize the total data center power consumption while satisfying both the QoS and the thermal constraints. The QoS constraint requires the workload execution rate of every zone to be greater than or equal to the workload arrival rate at the zone

$$\boldsymbol{\mu}(\tau) \geq diag\{\mathbf{1}\lambda^W(\tau)\}\boldsymbol{s}(\tau), \quad (10)$$

where $diag\{\boldsymbol{x}\}$ is the diagonal matrix having the elements of the vector $\boldsymbol{x}$ on the main column and $\mathbf{1}$ is the vector of appropriate dimension whose elements are all 1. A more general formulation of the computational requirements can be obtained by considering the profit obtained by executing the workload with a certain QoS in the controller's cost function. In such a case, the goal of the control becomes the search of the best tradeoff between minimizing the cost of powering the data center and maximizing the profit induced by executing the workload. Initial results in this direction can be found in [47].

The thermal constraints on IT devices are formulated in terms of upper bounds on the input temperature of the thermal nodes

$$\boldsymbol{T}_{\text{in}}(\tau) \leq \overline{\boldsymbol{T}_{\text{in}}}. \quad (11)$$

Controllable variables are also subject to constraints. Constraints on the vector of workload execution rates are given by

$$\mathbf{0} \leq \boldsymbol{\mu}(\tau) \leq \overline{\boldsymbol{\mu}}, \quad (12)$$

where the inequalities are applied componentwise. Controllers do not migrate workload among the zones, i.e.,

$$\boldsymbol{\delta}(\tau) = \mathbf{0}, \quad (13)$$

where $\mathbf{0}$ is the zero vector. The constraints on the vector of workload scheduling are given by

$$\mathbf{0} \leq \boldsymbol{s}(\tau) \leq \mathbf{1}, \qquad \mathbf{1}^T \boldsymbol{s}(\tau) \leq 1. \quad (14)$$

The second constraint in (14) allows the data center controller to drop workload. The constraints on the vector of reference temperatures are given by

$$\underline{\boldsymbol{T}_{\text{ref}}} \leq \boldsymbol{T}_{\text{ref}}(\tau) \leq \overline{\boldsymbol{T}_{\text{ref}}}. \quad (15)$$

Baseline controller. The baseline control approach is a reasonable control approach when the power consumed by the CT is much smaller than the total power consumed by IT. A drawback of this control approach is that it cannot guarantee that QoS and the thermal constraints are enforced.

For all $\tau \in \mathbb{R}$, the baseline controller sets the control variables to

$$\begin{aligned} \boldsymbol{\mu}(\tau) &= \overline{\boldsymbol{\mu}}, & \boldsymbol{\delta}(\tau) &= 0, \\ \boldsymbol{s}(\tau) &= \mathbf{1}\frac{1}{N}, & \boldsymbol{T}_{\text{ref}}(\tau) &= \underline{\boldsymbol{T}_{\text{ref}}}. \end{aligned} \quad (16)$$

Uncoordinated controller. The uncoordinated controller represents a control strategy typically found in modern data centers where the management of the IT and the CT are assigned to two uncoordinated controllers. The uncoordinated controller is obtained by the composition of two controllers. The first controller deals with the optimization of the IT. The second controller manages the CT. Both controllers consider a predictive, discrete-time model of the data center. In this case, the QoS and the thermal constraints are considered and enforced only at the beginning of every interval.

Let $\mathcal{T}$ be the time horizon considered by the two optimization problems. We define $\hat{\boldsymbol{\mu}}(h|k)$ as the predicted value of the variable $\boldsymbol{\mu}(\tau)$ at the beginning of the h^{th} interval based on the information available up to the beginning of the k^{th} interval and define the set $\mathcal{M} = \{\hat{\boldsymbol{\mu}}(k|k), \ldots, \hat{\boldsymbol{\mu}}(k+$

$\mathcal{T}-1|k)\}$. Similarly, we define the variables $\hat{\boldsymbol{s}}(h|k)$, $\hat{\boldsymbol{\delta}}(h|k)$, and $\hat{\boldsymbol{T}}_{\text{ref}}(h|k)$, and the sets $\mathcal{S} = \{\hat{\boldsymbol{s}}(k|k), \ldots, \hat{\boldsymbol{s}}(k+\mathcal{T}-1|k)\}$, $\mathcal{D} = \{\hat{\boldsymbol{\delta}}(k|k), \ldots, \hat{\boldsymbol{\delta}}(k+\mathcal{T}-1|k)\}$, and $\mathcal{T}_{\text{ref}} = \{\hat{\boldsymbol{T}}_{\text{ref}}(k|k), \ldots, \hat{\boldsymbol{T}}_{\text{ref}}(k+\mathcal{T}-1|k)\}$. The predicted value of the workload arrival rate at the data center during the h^{th} interval, based on the information available up to the k^{th} interval, is denoted with $\hat{\lambda}(h|k)$. With $\hat{\mathbf{p}}_{\mathcal{N}}(h|k)$ we denote the expected average power consumption of the zones during the h^{th} interval, based on the information available up to the k^{th} interval. With $\hat{\mathbf{p}}_{\mathcal{C}}(h|k)$ we denote the expected average power consumption of the CRAC units during the h^{th} interval, based on the information available up to the k^{th} interval. At the beginning of every interval, the first part of the uncoordinated controller solves the following optimization problem

$$
\begin{aligned}
\textbf{1.} \quad & \min_{\mathcal{M},\mathcal{S},\mathcal{D}} \sum_{h=k}^{k+\mathcal{T}-1} \mathbf{1}^T \hat{\mathbf{p}}_{\mathcal{N}}(h|k) \\
& \textbf{s.t.} \quad \text{for all } h = k, \ldots, k+\mathcal{T}-1 \\
& \qquad \text{computational dynamics} \\
& \qquad \hat{\boldsymbol{\mu}}(h|k) \geq diag\{\mathbf{1}\hat{\lambda}^W(h|k)\}\hat{\boldsymbol{s}}(h|k) \\
& \qquad \mathbf{0} \leq \hat{\boldsymbol{\mu}}(h|k) \leq \overline{\boldsymbol{\mu}}, \quad \hat{\boldsymbol{\delta}}(h|k) = \mathbf{0} \\
& \qquad \mathbf{0} \leq \hat{\boldsymbol{s}}(h|k) \leq \mathbf{1}, \quad \mathbf{1}^T \hat{\boldsymbol{s}}(h|k) \leq 1 \\
& \qquad \hat{\boldsymbol{l}}(k|k) = \boldsymbol{l}(k).
\end{aligned}
\tag{17}
$$

Based on the solution obtained for the optimization in (17), the second part of the uncoordinated controller generates and solves the following optimization problem

$$
\begin{aligned}
\textbf{2.} \quad & \min_{\mathcal{T}_{ref}} \sum_{h=k}^{k+\mathcal{T}-1} \mathbf{1}^T \hat{\mathbf{p}}_{\mathcal{C}}(h|k) \\
& \textbf{s.t.} \quad \text{for all } h = k, \ldots, k+\mathcal{T}-1 \\
& \qquad \text{thermal dynamics} \\
& \qquad \underline{\boldsymbol{T}_{\text{ref}}} \leq \hat{\boldsymbol{T}}_{\text{ref}}(h|k) \leq \overline{\boldsymbol{T}_{\text{ref}}} \\
& \qquad \hat{\boldsymbol{T}}_{\text{in}}(h+1|k) \leq \overline{\boldsymbol{T}_{\text{in}}} \\
& \qquad \hat{\boldsymbol{T}}_{\text{out}}(k|k) = \boldsymbol{T}_{\text{out}}(k)
\end{aligned}
\tag{18}
$$

Since the uncoordinated controller considers the computational and the thermal constraints in two different optimization problems, it cannot guarantee their enforcement. The uncoordinated controller manages variables related to both the cyber and the physical aspects of the data center and therefore, it is a cyber-physical controller. We call it uncoordinated because the management of the IT and CT are treated separately.

Coordinated controller. The coordinated control strategy is based on a discrete-time MPC approach and it manages the IT and the CT resources in a single optimization problem. Under mild assumptions, the coordinated controller is able to guarantee the enforcement of both the QoS and the thermal constraints [47]. The sets $\mathcal{M}$, $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{T}_{\text{ref}}$ are defined as in the uncoordinated controller case. At the beginning of every interval, the coordinated controller solves the following optimization problem

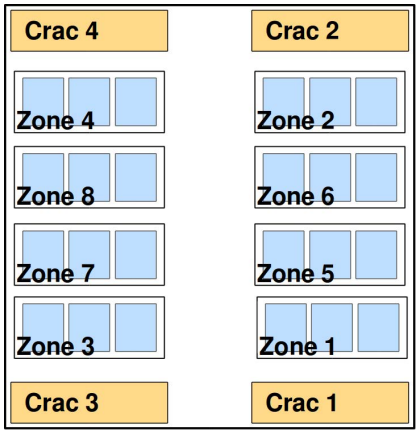


Fig. 5. An example of a data center layout. Blue rectangles represent groups of three racks each and yellow rectangles represent CRAC units.

$$
\begin{aligned}
& \min_{\mathcal{M},\mathcal{S},\mathcal{D},\mathcal{T}_{ref}} \sum_{h=k}^{k+\mathcal{T}-1} \mathbf{1}^T \hat{\mathbf{p}}_{\mathcal{N}}(h|k) + 1^T \hat{\mathbf{p}}_{\mathcal{C}}(h|k) \\
& \textbf{s.t.} \quad \text{for all } h = k, \ldots, k+\mathcal{T}-1 \\
& \qquad \text{computational dynamics} \\
& \qquad \text{thermal dynamics} \\
& \qquad \hat{\boldsymbol{\mu}}(h|k) \geq diag\{\mathbf{1}\hat{\lambda}^W(h|k)\}\hat{\boldsymbol{s}}(h|k) \\
& \qquad \mathbf{0} \leq \hat{\boldsymbol{\mu}}(h|k) \leq \overline{\boldsymbol{\mu}}, \qquad \hat{\boldsymbol{\delta}}(h|k) = \mathbf{0} \\
& \qquad \mathbf{0} \leq \hat{\boldsymbol{s}}(h|k) \leq \mathbf{1}, \qquad \mathbf{1}^T \hat{\boldsymbol{s}}(h|k) \leq 1 \\
& \qquad \underline{\boldsymbol{T}_{\text{ref}}} \leq \hat{\boldsymbol{T}}_{\text{ref}}(h|k) \leq \overline{\boldsymbol{T}_{\text{ref}}}, \quad \hat{\boldsymbol{T}}_{\text{in}}(h+1|k) \leq \overline{\boldsymbol{T}_{\text{ref}}} \\
& \qquad \hat{\boldsymbol{l}}(k|k) = \boldsymbol{l}(k), \qquad \hat{\boldsymbol{T}}_{\text{out}}(k|k) = \boldsymbol{T}_{\text{out}}(k)
\end{aligned}
\tag{19}
$$

A drawback of the coordinated controller is the complexity of the optimization problem that has to be solved. Typically the optimization problem is non-convex and large. Local optimal solutions may yield control strategies that are worse than the those obtained by an uncoordinated controller.

V. Simulation Results

We evaluate the long-term performance of the three control strategies for multiple constant workload arrival rates. The performance of the three controllers are evaluated in the ideal case, where the controllers have perfect knowledge about the data center and when the data center has reached its thermal and computational equilibrium. When the robustness of the control approaches is the focus of the simulation, then modeling errors and prediction errors should also be considered in the simulation. The simulations are developed using the TomSym language with Knitro as the numerical solver.[2]

We consider the data center layout depicted in Fig. 5. The picture represents a small data center containing 32 racks and 4 CRAC units. The CT comprises 4 CRAC units which cool the servers through a raised-floor architecture. Racks are grouped into eight zones and every rack contains 42 servers. Under the same workload, servers in zones 5 to 8 consume

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

TABLE I
AVERAGE AND STANDARD DEVIATION VALUES OF $\sum_j \psi_{i,j}$. THE COEFFICIENTS REFER TO THE GRAPHS IN FIG. 7 AND FIG. 8.

$j \rightarrow$	Zones $1-4$		Zones $5-8$		CRACs	
$\downarrow i$	Avg.	Std.	Avg.	Std.	Avg.	Std.
Zones $1-4$	0.04	2.6e-6	0.03	2.2e-6	0.93	4.8e-6
Zones $5-8$	0.05	9.9e-7	0.52	4.8e-5	0.43	4.8e-5
CRACs	0.63	2.0e-5	0.25	4.3e-5	0.12	2.3e-5

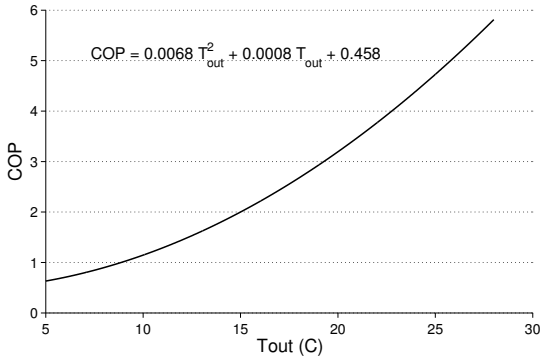


Fig. 6. Coefficient of performance of the CRAC units for different output temperature values [59].

10% less power than other servers, but they are not cooled as efficiently as the servers in zones 1 to 4 which are closer to the CRAC units. The maximum power consumption of the servers in zones 5 to 8 is 270 (W) and the maximum power consumption of the servers in zones 1 to 4 is 300 (W). It is assumed that every zone has a local controller that forces the zone to behave as postulated in the model, i.e., the amount of power consumed by every zone is proportional to the workload execution rate. CRAC units are identical and their efficiency, i.e., their COP, increases quadratically with respect to their output temperatures. The relationship between the output temperatures of the CRAC units and their COP is shown in Fig. 6. The coefficients relating input and output temperatures of zones and CRAC units are summarized in Table I. We set the maximum allowed input temperature of every zone at 27 °C, i.e., $\overline{\boldsymbol{T}_{\text{in}}} = 27$. This constraint reflects the environmental guidelines of the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) [60].

We define the the average utilization of the data center as the mean values of the ratios $\frac{\eta_i(\tau)}{\overline{\mu}_i}$ for $i = 1, \ldots, 8$. When the average utilization is 0 then zones process no workload, i.e., $\boldsymbol{\eta}(\tau) = \mathbf{0}$. When the average utilization is 1 then zones process the maximum amount of workload they can process, i.e., $\boldsymbol{\eta}(\tau) = \overline{\boldsymbol{\mu}}$. The behavior of the three controllers is not considered for small average utilization values since, at very low utilization values, nonlinearities of the IT and CT neglected in the proposed model become relevant. Fig. 7 shows the total data center power consumption obtained by the three controllers for different average utilization values.

The total data center power consumption obtained by the baseline controller grows proportionally with the average utilization. The proportional growth is due to two factors.

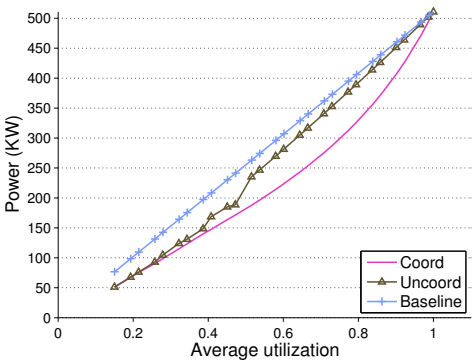


Fig. 7. Average data center power consumption for different utilization values.

The first factor is the assumption that rack-level and server-level controllers make the power consumption of every zone to grow proportionally with the amount of workload they process. The second factor is that the reference temperature of each CRAC unit is fixed and always lower than or equal to its input temperature. In such a case, the efficiency of every CRAC unit is constant and the power consumed by the CT grows proportionally with the amount of power that the zones consume.

The total data center power consumption obtained by the uncoordinated controller is always lower than the one obtained by the baseline controller. This happens because the uncoordinated controller assigns as much workload as possible to the most energy-efficient servers, i.e., those servers located in zones 5 to 8, and it tries to maximize the efficiency of the CT by setting the reference value of every CRAC unit to the largest value that still enforces the thermal constraints.

The additional savings obtained by the coordinated controller are due to the coordinated management of the IT and CT resources. In particular, the coordinated controller, depending on the amount of workload that has to be processed by the zones, decides whether it is more efficient to allocate workload to the energy-efficient servers, i.e., to the servers in zones 5 to 8, or to the efficiently cooled servers, i.e., to the servers in zones 1 to 4.

The PUE values obtained by the three controller are shown in Fig. 8. In the baseline controller case, the CRAC units operate at a constant, minimum, efficiency and therefore, the PUE values obtained by the baseline controller are constant. The uncoordinated and the coordinated controllers are able to improve, i.e., to lower, the PUE obtained by the baseline controller. The PUE curves obtained by the uncoordinated and by the coordinated controller are not smooth because for some values of the workload arrival rate, the controllers are unable to locate the true minimum of the optimization problems they solve. In these cases, control actions are based on a local minimum that may be different from the control actions chosen for nearby values of the workload arrival rate.

In the second simulation we focus on a case where the inlet temperatures of the servers in zones 1 to 4 equal the supplied air temperatures of the CRAC units, i.e., $\sum_j \psi_{i,j} \simeq 1$ when i is the index of the zones 1 to 4 and j represents

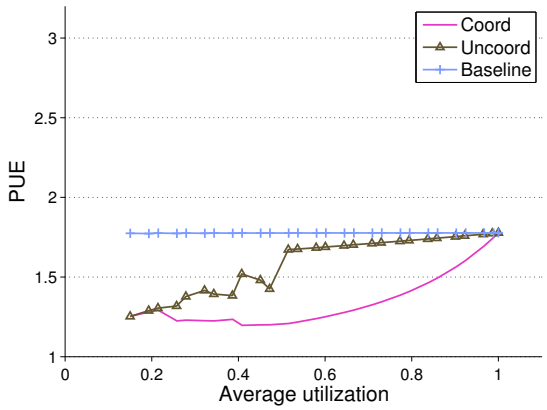


Fig. 8. PUE values obtained by the baseline, uncoordinated, and coordinated controller.

TABLE II
AVERAGE AND STANDARD DEVIATION VALUES OF $\sum_j \psi_{i,j}$. THE COEFFICIENTS REFER TO THE GRAPHS IN FIG. 9 AND FIG. 11.

$j \rightarrow$	Zones $1-4$		Zones $5-8$		CRACs	
$\downarrow i$	Avg.	Std.	Avg.	Std.	Avg.	Std.
Zones $1-4$	0	0	0	0	1	0
Zones $5-8$	0.3	2.9e-5	0.4	8.0e-6	0.30	2.9e-5
CRACs	0.51	5.6e-5	0.34	3.4e-5	0.15	2.5e-5

the CRAC units only. Also, in the second simulation the servers in zones 1 to 4 are subject to less air-recirculation than in the first simulation case and their inlet temperatures depends more on the output temperatures of the servers in zones 5 to 8. The total power consumption obtained by the uncoordinated equals the total power consumption obtained by the coordinated controller for every average utilization value, i.e., there is no loss of performance in managing IT and CT separately. Fig. 9 shows the total data center power consumption obtained by the three controllers in the second simulation. Table II summarizes the values of the coefficients relating input and output temperatures of zones and CRAC units for this simulation. The other parameters did not change.

The third simulation represents a data center where almost half of the airflow cooling servers in zones 5 to 8 comes from other servers, i.e., the third simulation considers a data center case where some servers are poorly positioned with respect to the CRAC units. Table III summarizes the values of the coefficients relating input and output temperatures of zones and CRAC units for this simulation. The other parameters did not change. Fig. 10 shows the total data center power consumption obtained by the three controllers in the third simulation.

The PUE obtained by the three controllers for these new cases are shown in Figs. 11 and 12. As shown in Fig. 12, when there exists a large variability among the server cooling efficiency, the PUE strongly depends on the average utilization of the data center.

TABLE III
AVERAGE AND STANDARD DEVIATION VALUES OF $\sum_j \psi_{i,j}$. THE COEFFICIENTS REFER TO THE GRAPHS IN FIG. 10 AND FIG. 12.

$j \rightarrow$	Zones $1-4$		Zones $5-8$		CRACs	
$\downarrow i$	Avg.	Std.	Avg.	Std.	Avg.	Std.
Zones $1-4$	0.08	0	0.08	0	0.84	0
Zones $5-8$	0.08	7.0e-8	0.66	4.8e-5	0.26	4.8e-5
CRACs	0.57	5.4e-5	0.18	9.0e-5	0.25	4.1e-5

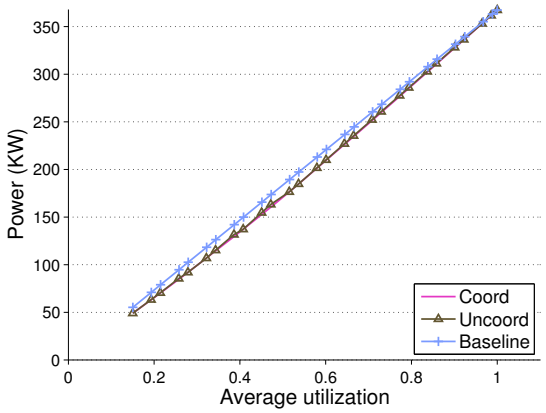


Fig. 9. Average data center power consumption for different utilization values. All of the zones are efficiently cooled.

VI. A CYBER-PHYSICAL INDEX FOR DATA CENTERS

For a given a data center, it would be useful to estimate a priori how much energy could be saved by using a coordinated controller rather than an uncoordinated one. Towards this end, we define *relative efficiency* as the ratio between area between the power consumption curve obtained by the uncoordinated controller and the power consumption curve obtained by the coordinated controller in Fig. 7, 9, and 10. With the appropriate weights, the relative efficiency can be mapped into the average monthly, or average yearly energy savings obtained by using a coordinated controller respect to an uncoordinated controller.

Consider a data center at its thermal and computational

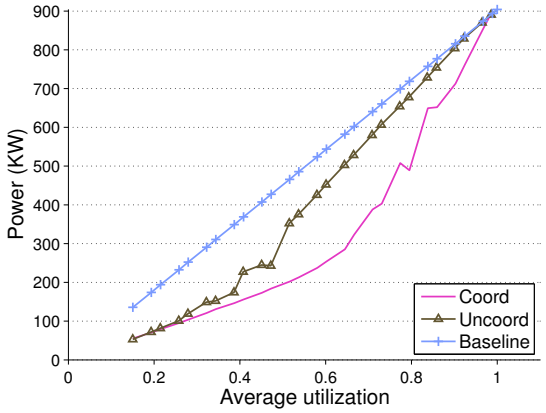


Fig. 10. Average data center power consumption for different utilization values. Large variability among cooling efficiency of the zones.

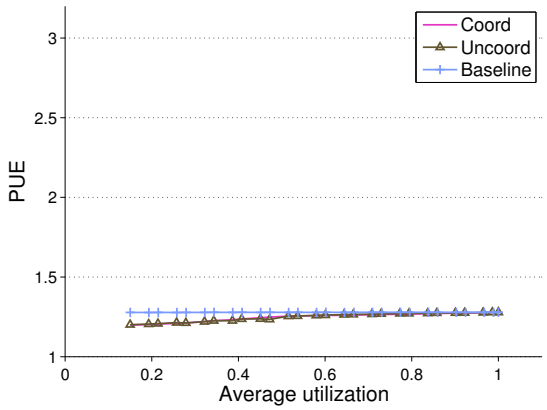


Fig. 11. PUE values obtained by the baseline, uncoordinated, and coordinated controller. All of the zones are efficiently cooled.

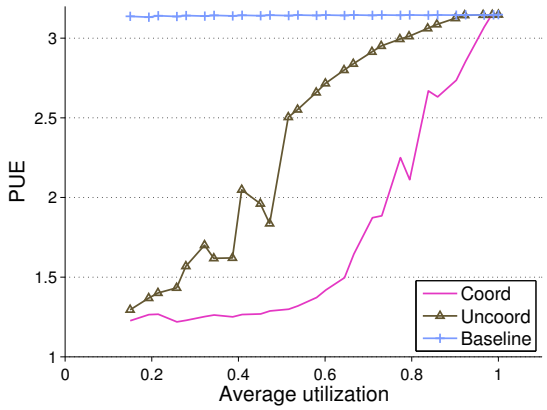


Fig. 12. PUE values obtained by the baseline, uncoordinated, and coordinated controller. Large variability among the cooling efficiency of the zones.

equilibrium and assume that both the QoS and the thermal constraints are satisfied. Furthermore, we assume that every CRAC unit provides a certain amount of cooling, i.e., for every thermal node j modeling a CRAC unit $T_{\text{ref},j} \leq T_{\text{in},j}$. Let i be the index of a cyber node representing a zone and let $T_{\text{in},i}$ be its input temperature value at thermal equilibrium. Collecting the input temperatures of the thermal nodes representing zones into vector $\boldsymbol{T}_{\text{in},\mathcal{N}}$, we define $\Psi_{[\mathcal{N},\mathcal{C}]}$ as the matrix composed of the $\{\psi_{i,j}\}$ variables such that i is the index of a thermal node modeling a zone and j is the index of a thermal node modeling a CRAC unit. We also collect all of the output temperatures of the thermal nodes modeling zones into the vector $\boldsymbol{T}_{\text{out},\mathcal{N}}$ and define $\Psi_{[\mathcal{N},\mathcal{N}]}$ as the matrix composed of the $\{\psi_{i,j}\}$ variables such that i and j are the indexes of two thermal nodes modeling zones. From the above, we can write

$$\boldsymbol{T}_{\text{in},\mathcal{N}} = \Psi_{[\mathcal{N},\mathcal{N}]}\boldsymbol{T}_{\text{out},\mathcal{N}} + \Psi_{[\mathcal{N},\mathcal{C}]}\boldsymbol{T}_{\text{ref}}. \tag{20}$$

With a slight abuse of notation, we use the symbol $diag\{\frac{\alpha_i \mathsf{c}_i}{\mathsf{k}_i}\}$ to denote the diagonal matrix composed of the $\{\frac{\alpha_i \mathsf{c}_i}{\mathsf{k}_i}\}$ terms, where k_i, c_i, and α_i are the coefficients introduced in (6) and in (7). Assuming every zone is processing a constant amount of workload and the matrix $(I - \Psi_{\mathcal{N},\mathcal{N}})$ is invertible, (20) can be rewritten as

$$\begin{aligned}\boldsymbol{T}_{\text{in},\mathcal{N}} &= (I - \Psi_{[\mathcal{N},\mathcal{N}]})^{-1} diag\Big\{\frac{\alpha_i \mathsf{c}_i}{\mathsf{k}_i}\Big\}\boldsymbol{\eta} + \Psi_{[\mathcal{N},\mathcal{C}]}\boldsymbol{T}_{\text{ref}} \\ &= \mathcal{L}\boldsymbol{\eta} + \Psi_{[\mathcal{N},\mathcal{C}]}\boldsymbol{T}_{\text{ref}},\end{aligned} \tag{21}$$

where $\mathcal{L} = (I - \Psi_{[\mathcal{N},\mathcal{N}]})^{-1} diag\{\frac{\alpha \mathsf{c}_i}{\mathsf{k}_i}\}$ and $\boldsymbol{\eta} = \Big[\eta_1, \ldots, \eta_N\Big]^T$ is the vector of workload departure rates from every zone at the equilibrium.

The variation of the input temperature of the thermal nodes with respect to a variation of the workload execution rate in the computational nodes, or to a variation of the reference temperature vector can be written as

$$\frac{\partial \boldsymbol{T}_{\text{in},\mathcal{N}}}{\partial \boldsymbol{\eta}} = \mathcal{L}, \qquad \frac{\partial \boldsymbol{T}_{\text{in},\mathcal{N}}}{\partial \boldsymbol{T}_{\text{ref}}} = \Psi_{[\mathcal{N},\mathcal{C}]}. \tag{22}$$

The physical meaning of the variables $\{\psi_{i,j}\}$ implies that the matrix $(I - \Psi_{\mathcal{N},\mathcal{N}})$ is invertible when the input temperatures of all the thermal nodes representing a zone are affected by at least one thermal node representing a CRAC unit.

The inlet temperature of an efficiently cooled server largely depends on the reference temperature of the CRAC units and marginally on the execution rate of other servers. Let i be the index of a thermal node representing a zone. The i^{th} node is efficiently cooled if [3]

$$\Big\|\frac{\partial T_{\text{in},i}}{\partial \boldsymbol{T}_{\text{ref}}}\Big\|_2 \gg \Big\|\frac{\partial T_{\text{in},i}}{\partial \boldsymbol{\eta}}\Big\|_2.$$

We consider the vector $\boldsymbol{z} = \Big[\boldsymbol{T}_{\text{ref}} \quad \boldsymbol{\eta}\Big]^T$ and define the *relative sensitivity index* of the i^{th} node as

$$\mathcal{S}_i = \Big\|\frac{\partial T_{\text{in},i}}{\partial \boldsymbol{T}_{\text{ref}}}\Big\|_2 \Big/ \Big\|\frac{\partial T_{\text{in},i}}{\partial \boldsymbol{z}}\Big\|_2.$$

When the relative sensitivity index of the i^{th} zone equals 1, the input temperature of the i^{th} zone uniquely depends on the reference temperature of the CRAC nodes, whereas when the relative sensitivity index equals 0, then the input temperature of the i^{th} zone only depends on the workload execution rate of the other zones.

A large variability among the relative sensitivity indexes can be exploited by a coordinated controller in order to improve the efficiency of a data center. We collect the relative sensitivity indexes in the vector $\boldsymbol{S}$ and define as *cyber-physical index* (CPI) of a data center the normalized standard variation of $\boldsymbol{S}$

$$CPI = k \cdot std(\boldsymbol{S}), \tag{23}$$

where k is the normalizing coefficient and std is the standard deviation of the elements of the vector argument.

Fig. 13 shows the relative efficiency obtained by the coordinated controller for different values of the CPI. When the CPI values are larger than about 0.55, the uncoordinated controller is unable to find a cooling strategy that satisfies the thermal constraints for large values of the average data center utilization. When the CPI is almost 0, the relative efficiency of an uncoordinated controller is almost 0. As the CPI increases, the efficiency of a coordinated controller grows exponentially

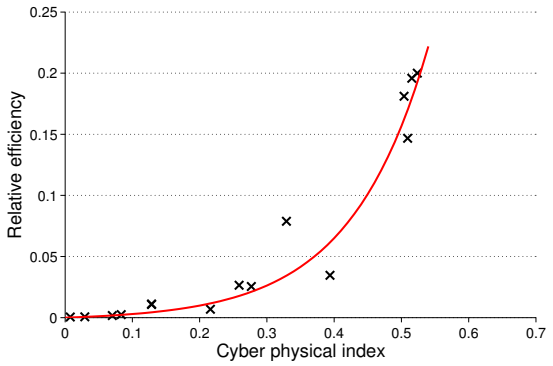


Fig. 13. Relative savings of the coordinated controller with respect to the uncoordinated controller.

fast. The simulation cases discussed in the previous section correspond to a CPI of 0.33, 0.04, and 0.52 respectively.

The exponential growth of the relative efficiency as a function of the CPI suggests that, in order to minimize the power consumption of a data center controlled by an uncoordinated controller, a careful positioning of the servers should be made. Different locations of a data center are subject to different cooling efficiency values. Given a data center, the relocation of some of its servers may move the CPI of the data center toward smaller values so that an uncoordinated controller will be as efficient as a coordinated controller.

VII. Discussion

This paper presents control of data centers from a CPS perspective. A survey of the literature and current practice shows how energy efficiency has been improved at all levels by taking into account the coupling between the cyber and physical features of data center components. We develop a control-oriented model of the coupled cyber and physical dynamics in data centers to study the potential impact of coordinating the control of the information technology with the control of the cooling technology at the data center level. Simulation results show that the the amount of savings that can be realized by coordinated control depends upon the amount of workload relative to the total data center capacity and the way the variations in the efficiency of servers are physically distributed relative to the physical distribution of cooling efficiency throughout the data center.

A new cyber-physical index (CPI) is proposed to quantify this dependence of the potential impact of coordinated control on the distribution of cyber (computational) and physical (power and thermal) efficiency. The CPI can be used to assess the need for coordinated control for a given data center, or as a tool to evaluate alternative data center designs. Further research is needed to understand how servers with different efficiencies should be distributed to reduce the need for coordinated control. We are also investigating improvements in the CPI to better quantify additional features, such as the impact of different efficiencies in CRAC units.

More research is needed to develop strategies for coordinating data center control with the power grid. Initial results in this direction can be found in [45], [47]. Data centers can play an important role in the smart grid because of their high power consumption density. A low-density data center can have a peak power consumption of 800 W/m^2 (75 W/sq.ft.), whereas a high-density data center can reach 1.6 KW/m^2 (150 W/sq.ft.) [12], [13], [61]. These values are much higher than residential loads, where the peak power consumption is about a few watts per squared meter [62], [63].

Finally, we believe the observations made in this paper concerning data centers from a CPS perspective can offer insights into how to understand and control other large-scale cyber-physical systems. Many cyber-physical systems can be viewed as coupled cyber and physical networks, similar to the computational and thermal networks used in this paper to model data centers. In CPS applications, it is important to understand the potential impact of coordinated control strategies versus uncoordinated strategies. Uncoordinated strategies offer the possibility of a "divide and conquer" approach to complexity, and in some cases the benefits of introducing more complex strategies to coordinate cyber and physical elements of a system may not be significant. The CPI defined for data centers offers one example of how to measure the potential impact of coordinated cyber and physical control. We expect that developing similar indices for other large-scale CPS applications could be of value.

[3]We focus on 2-norm, but other norms can be considered.

Acknowledgments

The authors would like to thank the research group at the Parallel Data Lab of Carnegie Mellon University and the researchers at the Sustainable Ecosystem Research Group of HP Labs for the helpful discussions and useful advice on modeling and control of data centers.

References

[1] R. C. Chu, R. E. Simons, M. J. Ellsworth, R. R. Schmidt, and V. Cozzolino, "Review of cooling technologies for computer products," *IEEE Transactions on Device and Materials Reliability*, vol. 4, no. 4, pp. 568–585, 2004.

[2] T. J. Breen, E. J. Walsh, J. Punch, A. J. Shah, and C. E. Bash, "From chip to cooling tower data center modeling: Part I influence of server inlet temperature and temperature rise across cabinet," in *Proc. of the 12th IEEE Intersociety Conf. Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM)*, Jun. 2010, pp. 1–10.

[3] E. J. Walsh, T. J. Breen, J. Punch, A. J. Shah, and C. E. Bash, "From chip to cooling tower data center modeling: Part II influence of chip temperature control philosophy," in *Proc. of the 12th IEEE Intersociety Conf. Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM)*, Jun. 2010, pp. 1–7.

[4] C. D. Patel, R. K. Sharma, C. Bash, and M. Beitelmal, "Thermal considerations in cooling large scale high compute density data centers," in *ITHERM*, May 2002, pp. 767–776.

[5] M. H. Beitelmal and C. D. Patel, "Model-based approach for optimizing a data center centralized cooling system," Hewlett Packard Laboratories, Tech. Rep. HPL-2006-67, Apr. 2006.

[6] M. K. Patterson and D. Fenwick, "The state of data center cooling," Intel Corporation, White Paper, Mar. 2008.

[7] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *International Symposium on Computer Architecture*, Jun. 2007, pp. 13–23.

[8] J. Hamilton, "Cost of power in large-scale data centers," http://perspectives.mvdirona.com, Nov. 2008.

[9] U.S. Environmental Protection Agency, "Report to congress on server and data center energy efficiency," ENERGY STAR Program, Tech. Rep., Aug. 2007.

[10] C. D. Patel and A. J. Shah, "Cost model for planning, development and operation of a data center," Internet Systems and Storage Laboratory, HP Laboratories Palo Alto, Tech. Rep., Jun. 2005.

[11] S. Rahman, "Power for the internet," *Computer Applications in Power, IEEE*, vol. 14, no. 4, pp. 8–10, 2001.

[12] M. Patterson, D. Costello, P. F. Grimm, and M. Loeffler, "Data center TCO; a comparison of high-density and low-density spaces," Intel Corporation, White Paper, 2007.

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

[14] Hewlett-Packard, "Hp modular cooling system: water cooling technology for high-density server installations," Hewlett-Packard, Tech. Rep., 2007.

[15] J. Scaramella, "Worldwide server power and cooling expense 2006-2010 forecast," International Data Corporation (IDC), Sep. 2006.

[16] The Green Grid, "The green grid data center power efficiency metrics: PUE and DCiE," Technical Committee, White Paper, 2007.

[17] X. Zhu, D. Young, B. Watson, Z. Wang, J. Rolia, S. Singhal, B. McKee, C. Hyser, D. Gmach, R. Gardner, T. Christian, and L. Cherkasova, "1000 islands: Integrated capacity and workload management for the next generation data center," in *International Conference on Autonomic Computing*, Jun. 2008, pp. 172–181.

[18] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, and X. Zhu, "No "power" struggles: Coordinated multi-level power management for the data center," in *Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, Mar. 2008, pp. 48–59.

[19] Y. Cho and N. Chang, "Energy-aware clock-frequency assignment in microprocessors and memory devices for dynamic voltage scaling," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 6, pp. 1030–1040, 2007.

[20] H. Aydin and D. Zhu, "Reliability-aware energy management for periodic real-time tasks," *IEEE Transactions on Computers*, vol. 58, no. 10, pp. 1382–1397, 2009.

[21] P. Choudhary and D. Marculescu, "Power management of voltage/frequency island-based systems using hardware-based methods," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 3, pp. 427–438, 2009.

[22] J. Kim, S. Yoo, and C.-M. Kyung, "Program phase-aware dynamic voltage scaling under variable computational workload and memory stall environment," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 1, pp. 110–123, 2011.

[23] Z. Jian-Hui and Y. Chun-Xin, "Design and simulation of the cpu fan and heat sinks," *IEEE Transactions on Components and Packaging Technologies*, vol. 31, no. 4, pp. 890–903, 2008.

[24] A. Mutapcic, S. Boyd, S. Murali, D. Atienza, G. De Micheli, and R. Gupta, "Processor speed control with thermal constraints," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 56, no. 9, pp. 1994–2008, 2009.

[25] A. Gandhi, M. Harchol-Balter, R. Das, and C. Lefurgy, "Optimal power allocation in server farms," in *ACM SIGMETRICS*, Jun. 2009, pp. 157–168.

[26] A. Gandhi, M. Harchol-Balter, and I. Adan, "Server farms with setup costs," *Performance Evaluation*, vol. 67, pp. 1123–1138, 2010.

[27] M. Aghajani, L. Parolini, and B. Sinopoli, "Dynamic power allocation in server farms: a real time optimization approach," in *Proc. of the 49th IEEE Conference on Decision and Control*, Dec. 2010, pp. 3790–3795.

[28] A. Varma, B. Ganesh, M. Sen, S. R. Choudhury, L. Srinivasan, and B. Jacob, "A control-theoretic approach to dynamic voltage scheduling," in *International conference on compilers, architecture and synthesis for embedded systems*, Oct. 2003, pp. 255–266.

[29] J. Leverich, M. Monchiero, V. Talwar, P. Ranganathan, and C. Kozyrakis, "Power management of datacenter workloads using per-core power gating," *Computer Architecture Letters*, vol. 8, no. 2, pp. 48–51, 2009.

[30] R. Mahajan, C. pin Chiu, and G. Chrysler, "Cooling a microprocessor chip," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1476–1486, 2006.

[31] G. K. Thiruvathukal, K. Hinsen, K. Lufer, and J. Kaylor, "Virtualization for computational scientists," *Computing in Science & Engineering*, vol. 12, no. 4, pp. 52–61, 2010.

[32] P. Padala, X. Zhu, Z. Wang, S. Singhal, and K. G. Shin, "Performance evaluation of virtualization technologies for server consolidation," Hewelett-Packard, White Paper, 2007.

[33] N. Tolia, Z. Wang, P. Ranganathan, C. Bash, M. Marwah, and X. Zhu, "Unified power and cooling management in server enclosures," in *InterPACK*, Jul. 2009, pp. 721–730.

[34] X. Wang and Y. Wang, "Coordinating power control and performance management for virtualized server clusters," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, no. 2, pp. 245–259, 2011.

[35] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services," in *Proc. of the 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2008, pp. 337–350.

[36] H. Jin, L. Deng, S. Wu, X. Shi, and X. Pan, "Live virtual machine migration with adaptive, memory compression," in *Proc. IEEE Int. Conf. Cluster Computing and Workshops CLUSTER*, Aug. 2009, pp. 1–10.

[37] F. Ma, F. Liu, and Z. Liu, "Live virtual machine migration based on improved pre-copy approach," in *Proc. IEEE Int Software Engineering and Service Sciences (ICSESS) Conf*, 2010, pp. 230–233.

[38] P. Padala, K. G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, A. Merchant, and K. Salem, "Adaptive control of virtualized resources in utility computing environments," *SIGOPS Oper. Syst. Rev.*, vol. 41, pp. 289–302, Mar. 2007.

[39] J. B. Rawlings, "Tutorial overview of model predictive control," *Control Systems Magazine, IEEE*, vol. 20, no. 3, pp. 38–52, Jun. 2000.

[40] R. Scattolini, "Architectures for distributed and hierarchical model predictive control - A review," *Journal of Process Control*, vol. 19, no. 5, pp. 723–731, 2009.

[41] C. E. Bash, C. D. Patel, and R. K. Sharma, "Dynamic thermal management of air cooled data centers," in *Proc. of the 10th Intersociety Conf. Thermal and Thermomechanical Phenomena in Electronics Systems (ITHERM)*, no. 29, May. 2006, pp. 445–452.

[42] C. Bash and G. Forman, "Cool job allocation: Measuring the power savings of placing jobs at cooling-efficient locations in the data center," in *Proc. of USENIX Annual Technical Conference*, no. 29, Jun. 2007, pp. 363–368.

[43] E. Ferrer, C. Bonilla, C. Bash, and M. Batista, "Data center thermal zone mapping," Hewlett-Packard, White Paper, 2007.

[44] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *Proc. of the ACM SIGCOMM Conference on Data communication*, Aug. 2009, pp. 123–134.

[45] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment," in *Proc. of the 29th IEEE International Conference on Computer Communications (INFOCOM)*, Mar. 2010, pp. 1–9.

[46] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, "Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 11, pp. 1458–1472, 2008.

[47] L. Parolini, B. Sinopoli, and B. H. Krogh, "Model predictive control of data centers in the smart grid scenario," in *Proc. of the 18th International Federation of Automatic Control (IFAC) World Congress*, Aug. 2011, to appear.

[48] L. Parolini, N. Tolia, B. Sinopoli, and B. H. Krogh, "A cyber-physical systems approach to energy management in data centers," in *First international conference on cyber-physical systems*, Apr. 2010, pp. 168–177.

[49] M. Anderson, M. Buehner, P. Young, D. Hittle, C. Anderson, J. Tu, and D. Hodgson, "MIMO robust control for HVAC systems," *IEEE Transactions on Control Systems Technology*, vol. 16, no. 3, pp. 475–483, 2008.

[50] M. M. Toulouse, G. Doljac, V. P. Carey, and C. Bash, "Exploration of a potential-flow-based compact model of air-flow transport in data centers," in *American Society Of Mechanical Engineers ASME Conference*, Nov. 2009, pp. 41–50.

[51] Y. Chen, D. Gmach, C. Hyser, Z. Wang, C. Bash, C. Hoover, and S. Singhal, "Integrated management of application performance, power and cooling in data centers," in *Proc. of the 12th IEEE/IFIP Network Operations and Management Symposium (NOMS)*, Apr. 2010, pp. 615–622.

[52] L. A. Barroso and U. Holzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, 2007.

[53] A. Hawkins, "Unused servers survey results analysis," The Green Grid, White Paper 28, 2010.

[54] C. Lefurgy, X. Wang, and M. Ware, "Server-level power control," in *Proc. Fourth Int. Conf. Autonomic Computing ICAC*, Jun. 2007, p. 4.

[55] H. Kobayashi and B. L. Mark, *System Modeling and Analysis: Foundations of System Performance Evaluation*. Prentice Hall Press, 2008.

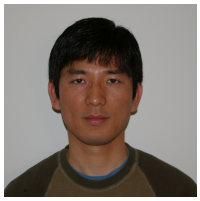
[56] L. Parolini, E. Garone, B. Sinopoli, and B. H. Krogh, "A hierarchical approach to energy management in data centers," in *Proc. of the 49th IEEE Conference on Decision and Control*, Dec. 2010, pp. 1065–1070.

[57] Q. Tang, T. Mukherjee, S. K. S. Gupta, and P. Cayton, "Sensor-based fast thermal evaluation model for energy efficient high-performance data centers," in *Fourth International Conference on Intelligent Sensing and Information Processing*, Oct. 2006, pp. 203–208.
[58] L. Parolini, B. Sinopoli, and B. H. Krogh, "Reducing data center energy consumption via coordinated cooling and load management," in *Workshop on Power Aware Computing and Systems (HotPower)*, Dec. 2008, pp. 14–18.
[59] J. Moore, J. Chase, P. Ranganathan, and R. Sharma, "Making scheduling "Cool": temperature-aware workload placement in data centers," in *USENIX Annual Technical Conference*, Apr. 2005, pp. 61–75.
[60] American Society of Heating, Refrigerating and Air-Conditioning Engineers, "Environmental guidelines for datacom equipment. expanding the recommended environmental envelope," ASHRAE, Tech. Rep., 2008.
[61] R. A. Greco, "High density data centers fraught with peril," EYP Mission Critical Facilities Inc., Slides, 2003.
[62] D. Meisegeier, M. Howes, D. King, and J. Hall, "Potential peak load reductions from residential energy efficient upgrades," ICF International, White Paper, 2002.
[63] National Association of Home Builders (NAHB) Research Center, Inc., "Review of residential electrical energy use data," NAHB Research Center, Inc., White Paper, 2001.

Luca Parolini received the B.Sc. degree in information engineering and the M.Sc. degree in automation engineering from the University of Padua, Padova, Italy, in 2004 and 2006, respectively. He is currently working towards the Ph.D. degree in electrical and computer engineering at Carnegie Mellon University, Pittsburgh, PA. His main research interests include analysis and control of data centers for energy efficiency and development of estimation and control techniques for information technology (IT) systems.

Bruno Sinopoli received the Dr. Eng. degree from the University of Padua, Padova, Italy, in 1998 and his M.S. and Ph.D. in Electrical Engineering from the University of California at Berkeley, in 2003 and 2005 respectively. After a postdoctoral position at Stanford University, Dr. Sinopoli joined the faculty at Carnegie Mellon University where he is an assistant professor in the Department of Electrical and Computer Engineering with courtesy appointments in Mechanical Engineering and in the Robotics Institute. Dr. Sinopoli was awarded the 2006 Eli Jury Award for outstanding research achievement in the areas of systems, communications, control and signal processing at U.C. Berkeley and the NSF Career award in 2010. His research interests include networked embedded control systems, distributed estimation and control over wireless sensor-actuator networks and cyber-physical systems security.

Bruce H. Krogh is professor of electrical and computer engineering at Carnegie Mellon University, Pittsburgh, PA. He received the BS degree in mathematics and physics from Wheaton College, Wheaton, IL, in 1975, and the MS and PhD degrees in electrical engineering from the University of Illinois, Urbana, in 1978 and 1982, respectively. He was a past Associate Editor of the IEEE Transactions on Automatic Control and Discrete Event Dynamic Systems: Theory and Applications, and founding Editor-in-Chief of the IEEE Transactions on Control Systems Technology. His current research interests include synthesis and verification of embedded control software, discrete event and hybrid dynamic systems, and distributed control strategies for the smart grid and other energy-related applications.



Zhikui Wang received the B.Eng. degree in process automation and automatic instrumentation from Tsinghua University, Beijing, China, in 1995, the M.Eng. degree in industrial automation from the Chinese Academy of Sciences, Beijing, China, in 1998, and the Ph.D. degree in electrical engineering, control systems from the University of California, Los Angeles in 2005.

He is a Senior Research Scientist at HP Laboratories, Hewlett-Packard Company, Palo Alto, CA, which he joined in 2005. His research interests are in application of control and optimization techniques in sustainable information technology (IT) ecosystems, including application performance control, workload management, data center power and cooling control.