Dynamic Workload and Cooling Management in High-Efficiency Data Centers

Marina Zapater · Ata Turk , José M. Moya , José L. Ayala , Ayse K.Coskun

Abstract—**Energy efficiency research in data centers has traditionally focused on raised-floor air-cooled facilities. As rack power density increases, traditional cooling is being replaced by close-coupled systems that provide enhanced airflow and cooling capacity. This work presents a complete model for close-coupled data centers with free cooling, and explores the power consumption trade-offs in these facilities as outdoor temperature changes throughout the year. Using this model, we propose a technique that jointly allocates workload and controls cooling in a power-efficient way. Our technique is tested with configuration parameters, power traces, and weather data collected from real-life data centers, and application profiles obtained from enterprise servers. Results show that our joint workload allocation and cooling policy provides 5% reduction in overall data center energy consumption, and up to 24% peak power reduction, leading to a 6% decrease in the electricity costs without affecting performance.**

I. Introduction

Energy efficiency in data centers continues to be an important research challenge. In 2010, data center electricity accounted for 1.3% of all the electricity use in the world [14], and this percentage has been growing since then. In year 2012, global data center power consumption increased to 38GW, with a further rise of 17% to 43GW in 2013 [24]. According to a 2014 report by the US Department of Energy, the top challenge for exascale research is energy efficiency [5].

Cooling power has traditionally been one of the major contributors to energy consumption in data centers, accounting for over 30% of the electricity bills in raised-floor air-cooled facilities [3]. A significant amount of research has been devoted to reduce cooling costs in these scenarios [1], [19], [27]. These solutions are mainly based on increasing room ambient temperatures when possible to reduce chiller power [17] and allocating the workload in a thermally-aware fashion [21].

Many cooling-aware techniques focus on reducing the Power Usage Effectiveness (PUE) of a data center, defined as the ratio between total facility power and IT power. According to a report by the Uptime Institute, average PUE reduced from 2.5 in 2007 to 1.65 in 2013 [17]. As chiller power is the most significant contributor of the overall non-IT power consumption, free cooling strategies have been implemented to achieve PUE values in the range of 1.1 to 1.3 [10].

Despite significant reductions in the amount of cooling power spent per server, data center power keeps increasing due to boosted rack power densities. Average maximum power density per rack increased from 9.3kW in 2013 to 11.7kW in 2014, and is projected to reach 50kW in 2015 [7]. Higher per-rack power densities also necessitate increased airflow and cooling capacities. These needs are addressed by replacing traditional Computer Room Air Conditioning (CRAC) systems with close-coupled cooling systems, such as in-row and in-rack cooling mechanisms, which bring cooling closer to the heat source [6]. In-row and in-rack cooling systems contain cold air supply and hot air exhaust inside a rack. This setup enables increased airflow to meet the demands of high-density racks and, at the same time, prevents hot air recirculation into the aisles, which helps lowering PUE.

As cooling becomes more efficient and the ratio of cooling power to overall power usage is reduced, savings obtained by increasing room temperatures becomes limited. In some cases, increasing room temperature may even lead to inefficiencies [18]. This is due to the impact of other contributors such as server fan power and leakage [26], which increase with high temperatures and may negate the savings achieved from higher room temperatures. Thus, trade-offs in IT and cooling power in such newer data centers need to be explored.

In this paper we propose the modeling and optimization of these newer type of energy-hungry, low-PUE data centers. We develop a workload allocation and cooling policy that is aware of application performance, energy requirements, and weather conditions. Our specific contributions are as follows:

- We provide a detailed model of high-density data centers with close-coupled cooling mechanisms (Section III). Our model considers the power consumption of chillers, towers, water pumps, in-rack coolers, and servers under both chiller operation and free cooling modes.
- We develop a cooling policy that sets per-rack inlet temperature and controls fan power according to the ambient temperature (Section IV). Our technique extends free cooling usage and also reduces the burden on chillers during warm weather.
- We propose a workload allocation policy that places jobs with similar power profiles in the same rack to balance per-rack temperature and increase cooling efficiency (Section V).

Proposed workload allocation and cooling management policies are able to work during runtime without incurring performance penalties. We evaluate our techniques using the configuration and workload traces of a real data center, and application profiles obtained from presently shipping enterprise servers. Results show that our workload and cooling management policy provides 5% reduction in the overall data center energy consumption, and 24% peak power reduction, with a 6% decrease in electricity costs due to both lower demand and energy costs.

II. Related Work

Energy optimization in data centers has been mostly focused on raised-floor air-cooled data centers [17]. In such systems, cooling efficiency is usually computed as a quadratic function of CRAC supply temperature [19], disregarding the impact of outside temperature on chiller power. Breen et al. [3] model the contributors to cooling separately (i.e., CRAC, chiller, tower and pumps), and highlight the adverse effects of increased inlet temperature in server leakage. However, they do not consider free cooling or newer cooling mechanisms. When using free cooling, overall cooling power is highly dependent on outdoor temperature, as shown in the recent work by Google [9]. However, this study disregards the effect of data room cooling control and workload allocation in energy consumption.

In data center rooms without hot-cold aisle containment, the hot air exhaust of servers usually recirculates to their inlet, and this recirculation generates non-uniform temperature profiles for the inlet temperature of servers within a rack. Because of the non-linear behavior of air and thermal dynamics, this process is modeled either via Computational Fluid Dynamics (CFD) simulations [16] or by linearized models that describe how much each server's outlet temperature affects other server inlet temperatures via a cross-interference matrix [23]. Several prior methods rely on these models to optimize cooling costs via workload allocation [19] or to reduce data center consumption via power budgeting [27].

As opposed to traditional cooling, close-coupled cooling systems achieve higher efficiency by increasing airflow and minimizing heat recirculation. In-row and in-rack coolers directly blow cold air to the inlet of servers within a rack and retrieve hot air from their outlets, eliminating recirculation and inlet temperature imbalances. A recent report by the Berkeley Labs [4] models the power consumption of these systems. However, their work uses a simple chiller model that disregards the effect of outdoor temperature and free cooling. Our work, on the contrary, shows the strong impact of these two factors on power consumption and cooling efficiency.

Kim et al. [13] characterize cooling power in a traditional CRAC-cooled data center via PUE, both in chiller and free cooling mode. They consider a fixed outdoor temperature threshold for free cooling usage, without considering the dependency between PUE and workload. Recent work by Schewedler [22] shows how outdoor temperature affects the properties of the cooling tower and the amount of heat that can be extracted from hot water, proving that the threshold for free cooling usage is not constant.

Unlike previous work, this paper explores the trade-offs and models the contributors to power in high-density close-couple cooling data centers with free cooling. For that purpose, we separately quantify the contributors to cooling power and computing power, both during chiller and free cooling operation, without relying on PUE estimation. We propose, for the first time, a joint workload and cooling control strategy to reduce overall data center power and cost for high-efficiency low-PUE data centers without performance degradation.

III. Modeling Low PUE Data Centers

In this section we describe our modeling methodology, which allows us to separate and quantify various contributors to cooling power. We also show some experimental observations about the trade-offs encountered in cooling-optimized high-density data centers, and use these results later to propose a joint workload and cooling management policy. To this end, we model the total power consumption of the data center P_{DC} that accounts for the sum of cooling power P_{cool} and computational power P_{IT}. We consider the different contributions to data center power as follows:

$$P_{DC} = P_{IT} + P_{cool}, \qquad P_{IT} = \sum_{k}^{server} P_{IT,k} \tag{1}$$

$$P_{IT,k} = P_{idle} + P_{leak,T} + P_{CPU,dyn} + P_{mem,dyn}, \tag{2}$$

$$P_{cool} = P_{fan} + P_{IRC} + P_{chiller} + P_{CT} + P_{pump}. \tag{3}$$

As server power is the greatest contributor to IT power consumption in a typical data center [8], we express IT power as the sum of the power of each server. This model can be extended to incorporate the contribution of storage systems and network switches. In Eq.(2), P_{idle} is server idle power, and $P_{CPU,dyn}$ and $P_{mem,dyn}$ are the dynamic power of CPU and memory, respectively. $P_{leak,T}$ is temperature-dependent leakage power, and increases exponentially with CPU temperature. These parameters have been modeled and experimentally validated in our previous work [26], by collecting power and temperature traces from a presently shipping highly multi-threaded SPARC-based enterprise server.

In Eq.(3), P_{fan} is the cumulative fan power of all servers, P_{IRC} represents the power consumption of in-row coolers, $P_{chiller}$ and P_{CT} are the chiller and cooling tower powers respectively, and P_{pump} is the power of the water pumps.

A. Cooling power modeling

Figure 1 shows the diagram of a data center in which a pod is cooled using a chiller and a tower. A pod is defined as two rows of racks containing high-density equipment with one dedicated in-row cooler (IRC) for every two racks. As opposed to traditional cooling in which a number of CRAC units pump cold air in the data room through the floor or ceiling, dedicated IRCs directly pump cold air inside the rack. This technique minimizes hot air recirculation. Thus, in a uniform inlet temperature profile to all servers inside a rack can be assumed [12]. Moreover, as racks are inclosed,

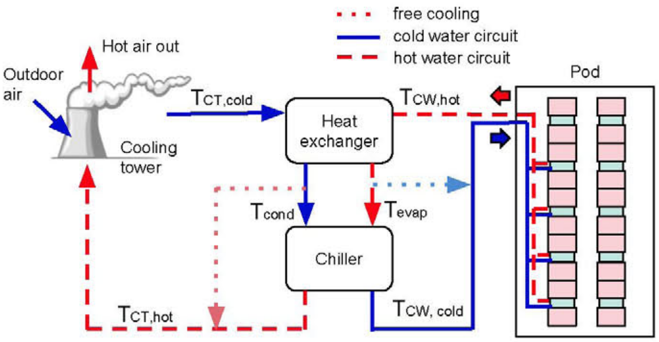

Fig. 1. Cooling model diagram. Dotted lines indicate free cooling mode.

TABLE I
PARAMETERS INVOLVED IN DATA CENTER MODELING

Parameter	Description	Units
T_{inlet}	Inlet temperature of servers	$^\circ C$
T_{outlet}	Outlet temperature of servers	$^\circ C$
$f_{air,fan}$	Server airflow	cfm
ρ_{air}, ρ_w	Air density, water density	kg/m^3
c_{air}, c_w	Specific heat capacity of air, water	kJ/(kg$^\circ C$)
T_{CWcold}	Cold water temperature received by IRCs	$^\circ C$
T_{CWhot}	Heated water returning from IRCs	$^\circ C$
$f_{IRCwater}$	Water flow through IRCs	gpm
T_{out}	Outdoor temperature	$^\circ C$
Q_{evap}	Heat transfer rate at the evaporator	W
T_{CThot}	Cooling tower hot water temperature	$^\circ C$
$T_{approach}$	Cooling tower approach temperature	$^\circ C$
T_{CTcold}	Cooling tower cold water temperature	$^\circ C$
λ_{CT}	Cooling tower load (%)	–
Q_{CT}	Heat transfer rate at the cooling tower	W

a different inlet temperature can be assigned to each rack. Therefore, the control knobs considered in our system are: i) per-IRC inlet temperature, ii) the workload executed, which affects server power consumption P_{IT}, and iii) server cooling, i.e., fan speed. The remaining parameters (such as IRC water flow) vary driven by the control knobs.

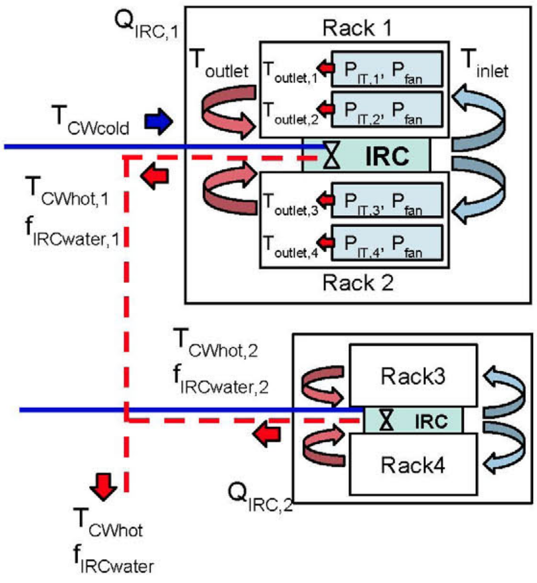

Fig. 2. Diagram of in-row cooler heat transfer

TABLE II
RELATION BETWEEN PARAMETERS IN DATA CENTER MODELING

	Description	Equation
(a)	Outlet temperature	$T_{outlet} = T_{inlet} + \frac{P_{IT}}{f_{fan,air} \cdot \rho_{air} \cdot c_{air}}$
(b)	IRC heat	$Q_{IRC,i} = \sum_n^{server} P_{IT,n}$
(c)	IRC hot water	$T_{CWhot,i} = T_{CWcold,i} + \frac{Q_{IRC,i}}{\rho_w \cdot c_w \cdot f_{IRCwater,i}}$
(d)	Total waterflow	$f_{IRCwater} = \sum(f_{IRCwater,i})$
(e)	Chiller hot water	$T_{CWhot} = \frac{\sum_i (f_{IRCwater,i} T_{cwhot,i})}{\sum_i f_{IRCwater,i}}$
(f)	Tower cold water	$T_{CTcold} = T_{out} + T_{approach}(T_{out}, \lambda_{CT})$
(g)	Evaporator heat	$Q_{evap} = \sum_i Q_{IRC,i}$
(h)	Cooling tower heat	$Q_{CT} = Q_{evap} + P_{chiller}$
(i)	Tower hot water	$T_{CThot} = T_{CTcold} + \frac{Q_{CT}}{\rho_w \cdot c_w \cdot f_{IRCwater}}$

1) Data center room cooling: IRCs are mainly composed of fans that transfer the heat generated by servers to water. Thus, their power consumption (P_{IRC}) increases in cubic relation with airflow, and can be modeled by directly measuring power and airflow, as shown in previous work [20].

Table I summarizes the main parameters involved in data center cooling. According to the laws of heat, the power consumed by servers P_{IT} causes an increase from inlet temperature (T_{inlet}) to server air exhaust ($T_{outlet,i}$), driven by server airflow ($f_{fan,air}$), as shown in Table II(a). IRC fans transfer all the heat generated in the servers they cool ($Q_{IRC,i}$, see Table II(b)) to the cold water coming from the chiller and tower. The fan speed of the IRCs needs to match the airflow of servers to avoid pressure imbalances. The performance curves of the IRC provide the waterflow needed ($f_{IRCwater,i}$) to transfer the generated heat to the water for a particular airflow. As a result, the cold water received by IRCs at temperature T_{CWcold} heats to a certain $T_{CWhot,i}$ (see Table II(c)). Figure 2 shows a diagram representing the air-water heat exchange performed at the IRC.

2) Chiller, tower, and pumps: After cooling down servers, water coming from all the IRCs mix, generating a total waterflow ($f_{IRCwater}$, Table II(d)) at a certain hot water temperature (T_{CWhot}, Table II(e)). The heat exchanger, chiller, and tower extract heat from water, supplying water back to the IRCs at T_{CWcold}. Table II(f-i) describes the main equations involved in the cooling infrastructure, which we explain next.

As shown in Figure 1, during free cooling, outdoor temperature (T_{out}) is low enough to bring hot water down to T_{CWcold} by means of the liquid-air heat exchange in the cooling tower, bypassing the chiller. In this scenario $P_{chiller} = 0$, but the remaining contributors to P_{cool} are still relevant. Free cooling is limited by two factors: i) the *range*, and ii) the *approach*. *Range* is defined as the maximum temperature difference between water entering and exiting the cooling tower ($T_{range} = T_{CThot} - T_{CTcold}$). *Approach* is the minimum difference between the cold water temperature (T_{CTcold}) and the outdoor temperature (T_{out}). It varies with outdoor temperature and tower load (see Table II(f)), due to the thermodynamic properties of air. At low outdoor temperatures the enthalpy of

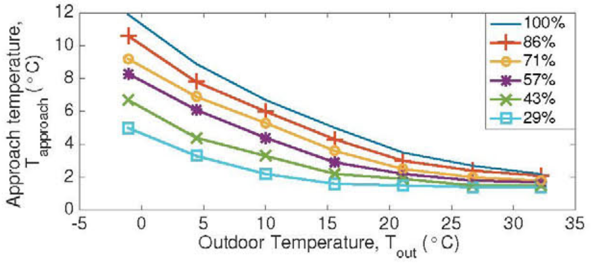

Fig. 3. Cooling tower approach temperature as a function of outdoor temperature for various tower loads ranging from 29% to 100%

air decreases. Therefore, air cannot hold as much moisture and approach temperature increases. This variation is documented in the tower manufacturer's datasheet. As IRCs need to receive cold water at a fixed temperature, the approach limits the usage of free cooling. Figure 3 shows the approach temperatures (taken from prior work [22]) for the cooling tower that are used throughout this paper.

During free cooling operation, chillers are turned off, drastically reducing cooling power. However, when outdoor temperature is above T_{CWcold}, the chiller carries the burden of removing heat from water. The power consumed during this process can be described as in Eq.(4):

$$P_{chiller} = \frac{Q_{evap}}{COP_{chiller}}, \tag{4}$$

where Q_{evap} is the heat load in the chiller evaporator side, and matches the heat generated by servers. Because servers are arranged in racks and a certain number of racks are cooled by each IRC, we can also express Q_{evap} as the sum of the heat generated by the servers cooled by each IRC (see Table II(g)). $COP_{chiller}$ is the coefficient of performance of the chiller [3]:

$$COP_{chiller} = \frac{T_{evap}}{T_{cond} - T_{evap}}, \tag{5}$$

where T_{evap} is the water temperature at the evaporator side of the chiller and T_{cond} is the temperature at the condenser side. The heat extracted from the hot water exiting IRCs is transferred to the cooling tower circuit. Moreover, the heat dissipated by chillers also needs to be extracted. Thus, the amount of heat that the tower needs to extract is the sum of both contributions (Table II(h)). This heat increases the tower water supply temperature to T_{CThot}.

The cooling tower, which is composed of fans, removes the heat from water using the outside air. We obtain the power consumption of the tower fans (P_{CT}) by fitting a third order polynomial using manufacturer's datasheets. The power consumption of water pumps P_{pump} is proportional to waterflow $f_{IRCwater}$ (obtained as in Table II(d)) and pressure drop ΔP, and inversely proportional to pump efficiency ϵ_{pump}:

$$P_{pump} = \frac{f_{IRCwater} \cdot \Delta P}{\epsilon_{pump}} \tag{6}$$

We consider $\Delta P = 32.4psid$ and $\epsilon_{pump} = 0.65$, as these are the values defined by ASHRAE for chilled water plant design, and are commonly used in commercial pumps [4].

B. Model validation

To calibrate and validate our developed models, we use data collected at the Massachusetts Green High Performance Computing Center[1] (MGHPCC). We utilize one month of correlated traces of cooling and IT power consumed by the MGHPCC facility in May 2014. In these traces, individual power contributions of chiller, tower, and pumps are available. The month of May is particularly interesting for model validation because, due to outdoor temperature variations (from 4°C to 24°C), the performance of both chiller and tower can be evaluated under various conditions. Moreover, we look at traces of per-rack IT power consumption, waterflow, and fan speed of each IRC in one pod. We also consider the corresponding outdoor temperature traces from May 2014. To tune and validate the models, we split the traces into two subsets: a first subset for model training and tuning (from May 1^{st} to May 15^{th}), and a second subset for testing (May 15^{th} to May 31^{st}).

The MGHPCC facility is equipped with 33 pods, each with 20 to 24 racks cooled down with IRCs. The hot water exhaust coming from the IRCs is cooled using three chillers and four cooling towers that extract the heat from the water leaving the IRCs. All IRCs have the same inlet temperature setting. A total of 10 pumps are used.

There are 8 to 12 IRCs per pod[2]. Five airflow and waterflow options are available in the IRCs. The IRCs have been characterized in previous work, both in terms of their heat exchange performance [4] and power consumption [20]. We use real IRC power measurements from MGHPCC to fit a third degree polynomial and obtain a model for IRC fan power. This fitted model exhibits 1.65% average error. We use the IRC performance curves to compute the hot water temperature exiting the IRC ($T_{CW,hot}$), given the waterflow, cold water temperature and heat load, obtaining a maximum error below $2^\circ C$ between our model and the MGHPCC data.

To validate the tower, chiller, and pump models, we use the overall IT power consumption of the MGHPCC facility. Given the waterflow of the IRCs of each pod, we compute overall waterflow in the data center, calculate pump power using our models, and compare our results to the pump power measurements at the MGHPCC. We obtain an average error of 2.63%. We use overall IT power consumption of the MGHPCC facility, hot water temperature and outdoor temperature to obtain tower power as a function of airflow. As expected, we observe that tower power is dependent on the dissipated heat (i.e., dependent on T_{CThot}). Finally, to validate chiller model, we use outdoor temperature to compute COP and chiller power using our models. We obtain COP values that are within the 2 to 8 range (i.e., the values expected in state-of-the-art chillers), and observe an increase in the chiller power through the month of May, as a consequence of the increase in outdoor temperature.

[1] http://www.mghpcc.org

[2] http://www.apc.com/products/resource/include/techspec_index.cfm?base_sku=ACRC500&tab=documentation

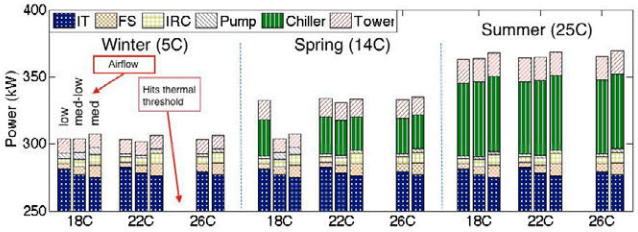

Fig. 4. Contributors to total data center power for various outdoor/inlet temperatures and airflow when all servers are fully utilized. Note that the y-axis minimum is 250kW.

C. Trade-offs in low-PUE data centers

To show the trade-offs in low-PUE data centers and evaluate the contribution of each component to overall power, we apply our models to a simulated data center that resembles one pod of the MGHPCC. We simulate a data center with 20 racks, each containing 20 servers with two 16-core SPARC T3 processors. Recall that we model IT power using the same technique in prior work [26], obtaining experimentally validated models for P_{fan}, $P_{leak,T}$, $P_{CPU,dyn}$, and $P_{mem,dyn}$.

The pod configuration assumed yields a maximum per-rack power of 16kW, which matches the average power limit of MGHPCC racks. To cool down the servers we assume the same IRC and chilled water parameters as in the MGHPCC (i.e., a maximum $T_{CWcold} = 18.3°C$ under free cooling).

As we are only simulating one pod, we need to scale the cooling infrastructure accordingly. For that purpose, we set the cooling and chiller assumptions following the methodology by Beitelmal et al. [2]. To extract all generated heat, we use a cooling tower with a range of 5°C and an approach of 2.5°C at a design wet-bulb temperature of 25.5°C, with a power consumption of $0.2hp/ton$ under maximum load. We use chillers with COP values that vary between 2 and 8 for low and high loads respectively. All values are selected from state-of-the-art chillers and towers [22] that match as closely as possible the configuration in the MGHPCC.

To explore the trade-offs in this scenario, we simulate power consumption when all servers are fully utilized, running a CPU-intensive application belonging to the SPEC CPU 2006 suite [11] for three different outdoor temperatures: $T_{out,winter} = 5°C, T_{out,summer} = 25°C, T_{out,spring} = 14°C$ under various rack inlet temperatures $T_{inlet} = \{18, 22, 26\}°C$ and airflows $f_{IRC,air} = \{low, med-low, high\}$.

Figure 4 shows the contributors to power consumption for each outdoor temperature. Within each season, we show results for three inlet temperatures, and for each inlet temperature we show three airflows. The setups corresponding to low airflow and 26°C inlet temperature always hit a CPU thermal threshold and, thus, are not considered valid solutions.

As seen in Figure 4, during winter, due to the usage of free cooling, chiller power is zero. PUE is around 1.08, and changing inlet temperature has a limited effect (at most 2.1% savings) on total power. The impact of temperature-dependent leakage on overall power is higher and letting inlet temperature raise to 26°C increases power compared to 22°C. Going below 22°C, however, requires increased waterflow through the IRC, resulting in inefficiencies. During summer, PUE varies from 1.26 to 1.52 depending on inlet temperature. In this case, the chiller dominates power consumption, and increasing rack inlet temperature reduces chiller effort. When outdoor temperature is similar to T_{CWcold}, i.e., $T_{out} = 14°C$, we approach the switching point between free cooling and chiller mode. By correctly setting inlet temperature and server fan speed, in this particular workload setup, we can extend the usage of free cooling without performance degradation, saving up to 16% power.

The best inlet and fan speed combination is workload-dependent, as workloads with high CPU power result in higher temperatures and higher leakage than memory-bound workloads [26]. Thus, they benefit from lower inlet temperatures both in summer and winter. Based on these observations, we can achieve substantial savings by appropriately tuning inlet and fan speed of servers. We see that setting a fixed ambient temperature throughout the year is not efficient. Moreover, different power profiles benefit from different inlet temperatures. Even though the amount of savings is dependent on the particular scenario, the trade-offs encountered and the conclusions obtained are valid for other servers and data center configurations using close-coupled cooling techniques.

IV. Cooling Control Policy

Our cooling policy uses the proposed models to set per-rack inlet temperature and airflow in to minimize data center power.

Algorithm 1 describes our policy in detail. Given the current outdoor temperature (T_{out}), the heat in each IRC ($Q_{IRC,i}$), and the cold water entering the room (T_{CWcold}), the policy first computes if outdoor temperature is sufficiently low to use free cooling (line 1), by calculating the approach temperature ($T_{approach}$) under various load conditions. If a cooling tower load (λ_{CT}) exists under which we can decrease water temperature to T_{CWcold}, then we are able to use free cooling.

Algorithm 1 Cooling management policy

```
Require: $T_{out}, Q_{IRC,i}, T_{cw,cold}$
 1: if $\exists\ \lambda_{CT} : T_{out} < (T_{cw,cold} - T_{approach}(T_{out}, \lambda_{CT}))$ then
 2:     free cooling: $T_{CThot} = T_{CWhot}$
 3:     budget $T_{CWhot,budget,i} \propto Q_{IRC,i}$
 4:     for all IRC do
 5:         for all $T_{inlet}$,$FS$ do
 6:             if $T_{CWhot,i} < T_{CWhot,budget,i}$ then
 7:                 isCandidate
 8:     if isCandidate $< 0$ then
 9:         select $IRC \leftarrow \min(P_{IT} + P_{fan} + P_{IRC,i})$
10: else
11:     $T_{evap,min} \leftarrow \max(COP)$
12:     budget $T_{evap,budget,i} \propto Q_{IRC,i}$
13:     for all IRC do
14:         $T_{CWhot,i} \leftarrow$sort($T_{inlet}$,$FS$)
15:         select $IRC_i \leftarrow min(T_{CWhot,i} > T_{evap,budget,i})$
16:         if $IRC_i = \emptyset$ then
17:             select $IRC_i \leftarrow max(T_{CWhot,i})$
```

Then, we compute the maximum tower hot water return temperature to use free cooling ($T_{CThot,MAX}$) and the maximum hot water temperature exiting the data room. This temperature limits the maximum heat that can be extracted. We budget the maximum per-IRC hot water temperature, proportionally to the generated heat ($Q_{IRC,i}$) (line 3), obtaining $T_{CWhot,budget,i}$. For each IRC, we exhaustively search the inlet temperature - fan speed pairs ($T_{inlet,i}, FS_i$) that do not exceed the maximum budget (lines 4-7). As the number of inlet and fan speed pairs is low (i.e., five different different fan speeds and inlet temperature pairs in our experiments), this search has low computational cost. Among the pairs that meet the budget, we use the one that minimizes the sum of IT, fan and IRC power (line 9). If no available pairs are found, then the chiller needs to be turned on.

The chiller is an important contributor to overall power. Thus, we aim to find the configuration that sets the minimum evaporator temperature ($T_{evap,min}$) to maximize COP. To this end, we compute $T_{evap,min}$ for the current outdoor conditions (line 11), according to Eq. 5. Then, we budget cooling among racks depending on their dissipated heat (line 12). For each IRC we sort all inlet and fan speed setups, according to $T_{CWhot,i}$ (line 14). We select the setting that exceeds the budget by the minimum amount (line 15). If no setup exceeds the minimum budget, we choose the one with highest evaporator temperature (lines 16-17).

V. Workload Allocation Policy

The goal of our workload allocation policy is to maximize the benefits of cooling control. We propose a *power-balance* policy, which places jobs with similar power profiles in servers sharing the same IRC. Such a grouping evens the cooling requirements of the servers sharing an IRC and enables our cooling control policy to set a fan speed and inlet temperature setup that avoids both over- and under-cooling.

We assume that incoming jobs are allocated one at a time and we assign each arriving job greedily to the racks that contain jobs with the highest power profile similarity to the arriving job. We characterize the power profiles of jobs via their dynamic CPU ($P_{cpu,dyn}$) and memory ($P_{mem,dyn}$) power consumption profiles [3]. Given these two values and applying leakage, cooling, and temperature models, the overall power consumption of a given job j_i running in a server can be computed [26]. Hence, the power profile $p(j_i) = p_i$ of a job j_i can be represented by the pair $(P^i_{cpu,dyn}, P^i_{mem,dyn})$ [4].

We consider workloads including independent jobs (a job may include multiple software threads or batches of tasks), each job requesting exclusive access to servers. Parameter-sweep type of applications that run the same program multiple times with different sets of parameters are good examples of these type of applications. This kind of jobs do not exhibit performance degradation due to locality issues, thus, our proposed policy implies no performance penalty.

Given the above assumptions, let $\mathcal{S}$=$\{s_1, \ldots, s_n\}$ be the set of servers in a data center, and let the servers be cooled by a set $\mathcal{C}$ of IRCs with server capacity κ. Without loss of generality, let $\mathcal{C}$=$\{C_1$=$\{s_1, \ldots, s_\kappa\}, \ldots, C_{n/\kappa}$=$\{s_{n-\kappa+1}, \ldots, s_n\}\}$ show the assignment of the set $\mathcal{S}$ of servers to the set $\mathcal{C}$ of IRCs in the data center. Let $\mathcal{J}^{cur}$=$\{j_1, \ldots, j_m\}$ be the set of jobs currently running, and let $\mathcal{A}^{cur} : \mathcal{S} \leftarrow \mathcal{J}^{cur}$ represent the current mapping of jobs to servers such that $\mathcal{A}^{cur}(s_k)$ is the job currently assigned to server s_k. Let t_i be the total time that job j_i requires to run, and t_i^{rem} the remaining time j_i needs to run in the server it is currently assigned to according to $\mathcal{A}^{cur}$. Note that $\mathcal{A}^{cur}(s_k) = \emptyset$ indicates that s_k is idle.

Given a job-to-server assignment $\mathcal{A}$, we can also define the power profile $p(s_j)$ of a server s_j as the power profile of job $\mathcal{A}(s_j)$. That is, $p(s_j)$=$p(\mathcal{A}(s_j))$. Note that if $\mathcal{A}(s_k)$=$\emptyset$, then $p(s_j) = (0, 0)$. Finally, we define the distance between the power profiles of two jobs as a linear combination of the distances between their respective CPU and memory profiles:

$$D(p_i, p_j) = \alpha|(P^i_{CPU,dyn} - P^j_{CPU,dyn})| + \\ (1-\alpha)|(P^i_{mem,dyn} - P^j_{mem,dyn})|, \quad (7)$$

where α is a scaling parameter tuned according to the relative effects of CPU and memory power on the cooling demand.

We propose to solve an iterative assignment problem, in which jobs are assigned one at a time. Given a current job-to-server assignment $\mathcal{A}^{cur} : \mathcal{S} \leftarrow \mathcal{J}^{cur}$, and a new job j^{new} to be run t^{new} seconds arrives to the system, we find the server assignment for j^{new} that minimizes the increase in the sum of distances between the power profiles of jobs already running in the same IRC after the assignment. Formally, we want to find the assignment that minimizes:

$$\sum_{s_j \in C_k \wedge s_\ell \in C_k \wedge \mathcal{A}^{cur}(s_j) = j_x \neq \emptyset} D(p^{new}, p(s_j)) min(t_x^{rem}, t^{new}). \quad (8)$$

Note that the distance of the newly assigned job to a previously assigned job in the same IRC is multiplied by the amount of time that both of them will be running.

The problem depends only on the currently running tasks and the incoming task to be assigned. It is straightforward to compute the optimization function presented in Eq.(8) for all possible assignment scenarios in an efficient manner and perform the greedy assignment as shown in Algorithm 2. The algorithm considers each IRC C_k that contains an idle server s_ℓ as assignment candidate and computes the distance of the power profile of t^{new} with all the active servers in C_k. Then, it multiplies the distance with the amount of time expected that these two jobs would be running in the same IRC if j^{new} were to be assigned to an idle server s_ℓ in C_k. The IRC with the minimum distance sum is selected and t^{new} is assigned to one of the idle servers in that IRC.

VI. Results

A. Experimental setup

To integrate the proposed modeling, workload allocation and cooling management strategies, and test our results with realistic traces of presently-shipping enterprise servers and data centers, our experimental setup uses three tools:

[3] We assume $P_{cpu,dyn}$ and $P_{mem,dyn}$ estimates are computed a priori for each job based on earlier runs

[4] As opposed to memory power, CPU power has a direct impact server maximum temperature and leakage. Therefore, it is not sufficient to match total dynamic power, and the profiles of CPU and memory need to be taken into account separately.

Algorithm 2 Power-balance ($\mathcal{A}^{cur} : \mathcal{S} \leftarrow \mathcal{J}^{cur}, j^{new}$)

```
Require: $\mathcal{C}, \mathcal{S}$
for each IRC $C_k \in \mathcal{C}$ do
    $CDist[k] \leftarrow 0$
for each IRC $C_k \in \mathcal{C}$ do
    if $\exists\, s_\ell \in C_k$ s.t. $\mathcal{A}^{cur}(s_\ell) = \emptyset$ then
        for each server $s_x \in C_k$ do
            if $\mathcal{A}^{cur}(s_x) \neq \emptyset$ then
                $t_t \leftarrow min(t^{new}, t_x^{rem})$
                $CDist[k] \leftarrow CDist[k] + D(p^{new}, p(s_x)) \times t_t$
$AssignedCooler \leftarrow indexOf(min(CDist))$
return $s_j \in AssignedCooler$ s.t. $\mathcal{A}^{cur}(s_j) = \emptyset$
```

- SLURM resource manager [25], an open-source tool that allocates exclusive and/or non-exclusive access to computer nodes so that users can perform work.
- An improved version of the SLURM-simulator tool [15], which allows the simulation of scheduling policies.
- Our custom designed data center power consumption simulation tool, *DCSim*. The simulator uses current outdoor temperature and workload assignment, and computes per-server power consumption, IRC, tower and chiller power, as well as overall data center power for a given cooling control policy. For this purpose, *DCSim* incorporates all the models described in Section III.

To integrate these tools, we developed a SLURM plugin responsible for workload allocation that calls our *power-balance* allocation policy and *DCSim* each time a job starts or ends. The parameters needed for job allocation (i.e., CPU/memory power profile, duration of the new jobs and current allocation) are passed to the allocation policy. The external allocator pauses the SLURM simulator, computes the allocation of the new task, and resumes the simulator execution. In the meantime, *DCSim* updates the power consumption values of each server, applies the cooling control policy, and computes the overall data center power consumption. This step includes computing per-server temperature, leakage and fan power.

B. Joint cooling and workload management

To test our workload allocation and cooling control policies we use the case study presented in Section III-C, i.e., a pod composed of 20 racks, each with 20 SPARC servers, and 1 IRC every 2 racks. We gather one year temperature traces from MGHPCC, in Holyoke, MA, USA.

For workload arrivals, we use SLURM arrival times and task duration of an HPC facility similar to MGHPCC running parameter-sweep applications. Namely, we use traces of the CEA-Leti supercomputer in France posted in the Parallel Workloads Archive [5]. These traces provide job arrival times and durations. To generate ($P_{CPU,dyn}, P_{mem,dyn}$) settings for each job in the trace, we run a set of SPEC CPU 2006 [11] applications, each with 4 to 256 simultaneous instances, on a real-life enterprise server (e.g., *mcf* x 4, *perlbench* x 128). We characterize a set of applications thar are diverse in their CPU and memory usage in order to obtain a wide range of possible power profiles. Then, for each job arrival in the trace, we randomly select a specific SPEC job and instance count (e.g., *mcf* x 4) and obtain the ($P_{CPU,dyn}, P_{mem,dyn}$) pair for that job from our database. This general methodology can be adapted to many other job or server types.

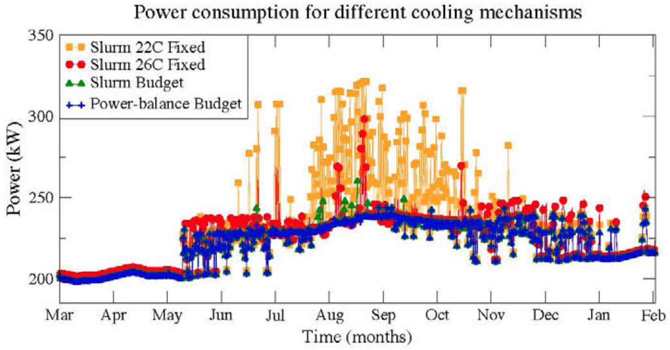

Fig. 5. Overall data center power consumption for various workload allocation and cooling management strategies

We test our proposed allocation policy against the default SLURM allocation, which selects a minimal number of consecutive nodes, in a round-robin fashion, to balance server usage. For each workload allocation, we test the following cooling control strategies:

- A fixed inlet temperature of 22°C for all IRC in the room throughout the year, which is the most common scenario in current data centers [17].
- A fixed 26°C that increases room temperature to minimize cooling power.
- *Budget*: our cooling-aware policy, that budgets cooling proportionally to rack heat.

Figure 5 shows the power consumption (in kW) for one year of execution of the investigated policies. As seen in the figure, fixed 26°C policy consumes less power than fixed 22°C during summer, due to the decreased power consumption of the chiller. However, its behavior is the opposite during spring and winter, because of the limited effect of increasing room temperature when free cooling is used, and the impact of server leakage and fan power. The SLURM Budget policy outperforms fixed ambient temperatures in reducing power, because it uses our models to predict power consumption and sets inlet temperatures to minimize energy. However, the SLURM allocation policy disregards workload characteristics, and can place workloads with very different power profiles in the same rack, causing overcooling. The *Power-balance Budget* policy uses our proposed joint workload and cooling policy to further reduce power consumption by avoiding cooling imbalances.

Table III shows the energy consumption, maximum power and electricity bill costs associated with each of the policies. Savings are computed against the SLURM allocation policy with a fixed inlet temperature of 22°C. As seen in the table, our *Power-balance Budget* strategy outperforms all other approaches, achieving savings in terms of energy, power, and economic costs. The savings in peak power consumption are particularly significant, reaching 24.2% savings for the *Power-balance Budget*. The savings in peak power are mainly due to a decreased burden on the chiller during the summer

[5] http://www.cs.huji.ac.il/labs/parallel/workload/l_cea_curie/

TABLE III
ENERGY, PEAK POWER AND ELECTRICITY COST FOR EACH POLICY

Policy	Energy	Energy savings(%)			Peak Power	Peak Power	Energy cost	Demand cost	Electric	Costs
	(KWh)	Yearly	Feb.	Aug.	(kW)	savings (%)	(thou. $)	(thou. $)	Bill (thou.$)	savings (%)
Slurm 22-fixed	2007	-	-	-	321.5	-	119.7	40.7	159.9	-
Slurm 26-fixed	1962	2.2	-0.4	10.9	298.2	-0.2	117.0	38.7	155.7	2.6
Slurm budget	1928	3.9	1.0	13.9	260.0	19.1	115.0	35.5	150.5	5.8
Power-balance 22-fixed	2033	-1.3	0.2	-3.4	321.7	-0.1	121.2	40.7	162.0	-1.3
Power-balance 26-fixed	1966	2.0	-0.4	10.9	322.1	-0.2	117.2	38.7	155.9	2.4
Power-balance budget	1926	4.1	1.2	14.4	243.8	24.2	114.9	34.9	149.8	6.3

months. Because the *Power-balance Budget* policy places workloads with similar power profiles together, distinguishing between CPU and memory consumption, the cooling effort can be focused on CPU intensive workload, which have higher leakage power than memory intensive workloads. The savings obtained in terms of peak power have an impact on the demand electricity cost of the data center, and increase computational capacity, enabling the deployment of more servers.

Our policy saves energy during all seasons in the year, showing more limited savings during winter (1.2%) and very high savings in summer (14.4%). The *Power-balance Budget* policy exploits the variability of workloads and outdoor conditions across time to reduce energy consumption, whereas the other policies disregard these factors. Electricity costs have been computed using the electricity rates of Holyoke [6]. As expected, the *Power-balance* policy, when not combined with cooling control, does not save energy, as it concentrates heat in certain racks without dedicating a proportional amount of cooling. If we scale the results obtained to the 33 pods in the MGHPCC, considering that all racks are occupied, and that the facility is running with an average utilization of 60%, savings would reach 170,000$/year.

VII. CONCLUSIONS

Despite recent advances in energy efficiency in data centers, unsustainable power consumption still represents an important challenge. In new highly-efficient data centers using close-coupled cooling mechanisms, other contributors to power consumption, such as fan or leakage power are becoming important. In this paper, we show how energy can be reduced by appropriately tuning cooling parameters in highly-efficient data centers. We develop an overall data center model that describes the relation between contributors to power, and propose a joint workload and cooling management strategy to set the per-rack inlet temperature and fan speed.

Our *Power-balance* allocation and *Budget* cooling control policy achieves 24% reduction in peak power and 5% reduction in data center energy. This leads to 6% savings in the yearly electricity bill without degrading performance.

REFERENCES

[1] A. Banerjee et al. Cooling-aware and thermal-aware workload placement for green hpc data centers. In *Proceedings of the International Conference on Green Computing*, pages 245–256, 2010.

[2] M. H. Beitelmal et al. Model-based approach for optimizing a data center centralized cooling system. Technical report, HP Labs, 2006.

[3] T. Breen et al. From chip to cooling tower data center modeling: Part I influence of server inlet temperature and temperature rise across cabinet. In *ITherm*, pages 1–10, 2010.

[4] H. Coles. Demonstration of rack-mounted computer equipment cooling solutions. Technical report, Berkeley National Laboratory, 2014.

[5] DOE ASCAC Subcommittee. Top ten exascale research challenges. Technical report, U.S. Department of Energy (DOE), february 2014.

[6] K. Dunlap and N. Rasmussen. Choosing between room, row, and rack-based cooling for data centers. *APC UK Co*, 2014.

[7] Emerson Electric, Inc. Data center 2025: Exploring the possibilities. Technical report, 2014.

[8] D. Floyer. Networks go green. http://wikibon.org/wiki/v/Networks_Go_GrEEN, 2011.

[9] J. Gao. Machine learning applications for data center optimization, 2014.

[10] Google Inc. Efficiency: How we do it. http://www.google.com/about/datacenters/efficiency/internal/.

[11] John L. Henning, SPEC CPU Subcommittee. SPEC CPU 2006 benchmark descriptions. http://www.spec.org/cpu2006/.

[12] K. B. John Niemann and V. Avelar. Impact of hot and cold aisle containment on data center temperature and efficiency. Technical report, Schneider Electric, 2013.

[13] J. Kim, M. Ruggiero, and D. Atienza. Free cooling-aware dynamic power management for green datacenters. HPCS, 2012.

[14] J. Koomey. Growth in data center electricity use 2005 to 2010. Technical report, Analytics Press, 2011.

[15] A. Lucero. Simulation of batch scheduling using real production-ready software tools. http://www.bsc.es/media/4856.pdf.

[16] L. Marshall and P. Bemis. Using CFD for data center design and analysis. Technical report, Applied Math Modeling White Paper, 2011.

[17] J. K. Matt Stansberry. Uptime institute 2013 data center industry survey. Technical report, Uptime Institute, 2013.

[18] R. L. Mitchell. Data center density hits the wall. http://www.computerworld.com/article/2522601/it-management/data-center-density-hits-the-wall.html, 2010.

[19] Moore et al. Making scheduling ”cool”: Temperature-aware workload placement in data centers. USENIX ’05.

[20] V. M. J. E. Rabassa, A. Economic performance of modularized hot-aisle contained datacenter pods utilizing horizontal airflow cooling, 2014.

[21] L. Ramos and R. Bianchini. C-oracle: Predictive thermal management for data centers. In *HPCA*, pages 111–122, 2008.

[22] M. Schwedler. Effect of heat rejection load and wet bulb on cooling tower performance. *ASHRAE Journal*, pages 16–23, January 2014.

[23] G. Varsamopoulos, A. Banerjee, and S. Gupta. Energy efficiency of thermal-aware job scheduling algorithms under various cooling models. In *Contemporary Computing*, volume 40. Springer, 2009.

[24] A. Venkatraman. Global census shows datacentre power demand grew 63% in 2012. http://www.computerweekly.com/news/2240164589/Datacentre-power-demand-grew-63-in-2012-Global\-datacentre-census, 2012.

[25] M. Yoo, A. B. Jette, M. A. Grondona. SLURM: Simple Linux Utility for Resource Management. *LNCS*, 2003.

[26] M. Zapater et al. Leakage-aware cooling management for improving server energy efficiency. *IEEE TPDS*, 2014.

[27] X. Zhan and S. Reda. Techniques for energy-efficient power budgeting in data centers. DAC, 2013.