J Supercomput
DOI 10.1007/s11227-010-0453-8

Proactive thermal management in green datacenters

Eun Kyung Lee · Indraneel Kulkarni · Dario Pompili · Manish Parashar

Abstract The increasing demand for faster computing and high storage capacity has resulted in an increase in energy consumption and heat generation in datacenters. Because of the increase in heat generation, cooling requirements have become a critical concern, both in terms of growing operating costs as well as their environmental and societal impacts. Presently, thermal management techniques make an effort to thermally profile and control datacenters' cooling equipment to increase their efficiency. In conventional thermal management techniques, cooling systems are triggered by the temperature crossing predefined thresholds. Such reactive approaches result in delayed response as the temperature may already be too high, which can result in performance degradation of hardware.

In this work, a proactive control approach is proposed that jointly optimizes the air conditioner compressor duty cycle and fan speed to prevent heat imbalance—the difference between the heat generated and extracted from a machine—thus minimizing the cost of cooling. The proposed proactive optimization framework has two objectives: (i) minimize the energy consumption of the cooling system, and (ii) minimize the risk of equipment damage due to overheating. Through thorough simulations comparing the proposed proactive heat-imbalance estimation-based approach against conventional reactive temperature-based schemes, the superiority of the proposed ap-

E.K. Lee (✉) · I. Kulkarni · D. Pompili · M. Parashar
NSF Center for Autonomic Computing, Department of Electrical and Computer Engineering, Rutgers University, 94 Brett Road, Piscataway, NJ 08854, USA
e-mail: eunkyung_lee@cac.rutgers.edu

I. Kulkarni
e-mail: indraneel_kulkarni@cac.rutgers.edu

D. Pompili
e-mail: pompili@cac.rutgers.edu

M. Parashar
e-mail: parashar@cac.rutgers.edu

Published online: 10 June 2010

proach is highlighted in terms of cooling energy, response time, and equipment failure risk.

Keywords Data center · Proactive approach · Modeling · Air cooling system · Thermal management

1 Introduction

The increasing demand for faster computing and high storage capacity has resulted in an increase in energy consumption and heat generation in datacenters. Because of the increase in heat generation, cooling requirements have become a critical concern, both in terms of growing operating costs as well as their environmental and societal impacts (e.g., increase in CO_2 emissions, overloading the electric supply grid resulting in power cuts, heavy water usage for cooling systems causing water scarcity, etc.) Many current datacenters are not following a sustainable model in terms of energy consumption growth as the rate at which computing resources are added exceeds the available and planned power capacities. For these reasons, there is a need for realizing environment friendly computing systems that maximize energy and cooling efficiency. Technical advances are leading to a pervasive computational ecosystem that integrates computing infrastructures with embedded sensors and actuators, thus giving rise to a new information/sensor-driven and autonomic paradigm for managing datacenter cooling systems.

Due to the increasing costs and high energy consumption of current cooling systems in datacenters, energy efficient and intelligent cooling solutions are required to minimize these costs and consumption. Empirical data from Little Blue Penguin cluster shows that every 10°C rise in temperature results in a doubling of the system failure rate, as per Arrhenius' equation applied to microelectronics, which increases the Total Cost of Ownership (TCO) significantly [11]. Overheating of components causes *thermal cycling*, which eventually leads to device failure, thus affecting the TCO [28]. Cooling systems aim at effectively maintaining the temperature of the datacenter. For robustness and safety, over-provisioning is often implemented to avoid any loss to property due to unforseen perturbations. According to Lawrence Livermore National Laboratory (LLNL), for every watt of power its IBM BlueGene/L consumes, 0.7 W is required to cool it [4, 14]. As the number of datacenters with high processing power increases, the expense to run cooling equipment such as chillers, compressors, and air handlers also increases. According to [6], it is predicted that datacenter energy consumption in the US will reach 100 billion kWh/year by 2011 with a corresponding energy bill of approximately $7.4 billion.

Current cooling solutions in datacenters rely on *reactive techniques*, which aim at keeping the temperature at a fixed value. Existing datacenter cooling systems control the temperature/humidity of the air based on the temperature external to the machines, i.e., the temperature inside a datacenter. Some of the cooling system control mechanisms are based on the internal temperature of the racks or blades. Irrespective of the external or internal temperature, the reactive approach has numerous disadvantages: (i) it takes a corrective action after the temperature has crossed a threshold and may

not be able to prevent damages in certain cases where temperature rises above the safe operating range of the internal components, (ii) it is very difficult to determine the optimal threshold range as it should not be too high that a small increase above it damages the components or too low to waste energy required for cooling, and (iii) if the threshold range is too small it causes cycling (or hysteresis) in the Computer Room Air-Conditioning (CRAC) unit and in turn reduces its life; conversely, if the threshold range is too high the response time of the system increases with a possible risk of damage to the internal components of the machines.

Due to the numerous disadvantages of temperature-based reactive approaches, in this work we propose a heat-imbalance estimation-based proactive approach that optimizes the cooling system operation by minimizing the cooling costs and the risk of damage to components due to overheating. The proposed proactive approach controls the cooling system before the heat imbalance can raise the temperature and cause damage to the internal components of a machine. Heat imbalance is the difference between the heat generated and heat extracted (under ideal conditions heat imbalance should be zero). The proactive approach has numerous advantages over the reactive approach: (i) there is no need for setting thresholds as needed in reactive approaches, (ii) it is intrinsically predictive in nature as it estimates the heat that will be generated in the future (based on information on scheduled type/intensity of workload) and, accordingly, adjusts the operation of the CRAC unit, (iii) it observes the "cause" instead of the "effect," i.e., it estimates the heat imbalance rather than measuring the rise in temperature caused by it.

The remainder of this article is organized as follows. In Sect. 2, we address the background and related work. In Sect. 3, we formulate our heat-imbalance estimation-based proactive control approach. In Sect. 4, we describe our mathematical model in details. In Sect. 5, we present the performance evaluation. Finally, in Sect. 6, we draw the conclusions and discuss future work.

2 Background and related work

There are two main approaches for thermal management of datacenters; one is *mechanical design based* and the other is *software based*. The former focuses on how to effectively distribute cold air by managing cooling infrastructure, while the latter focuses on how to balance or migrate jobs in such a way as to minimize heat imbalances.

Mechanical design-based approaches study the airflow models, datacenter design, and cooling system design. Datacenter design plays an important role in the efficient thermal management. The cooling systems used for current datacenters are chilled-water cooled CRAC units that supply a raised floor plenum underneath the racks with cold air. Perforated tiles are located near the racks to transfer the cool supply air to the front of the racks. The hot exhaust air from the racks is then collected from the upper section of the facility by the CRAC units, thus completing the airflow loop as shown in Fig. 1(a). Racks are typically arranged in rows with alternating airflow directions, forming "hot" and "cold" aisles [2]. This hot-and-cold aisle approach attempts to prevent mixing of the hot rack exhaust air and the cool supply air drawn into the

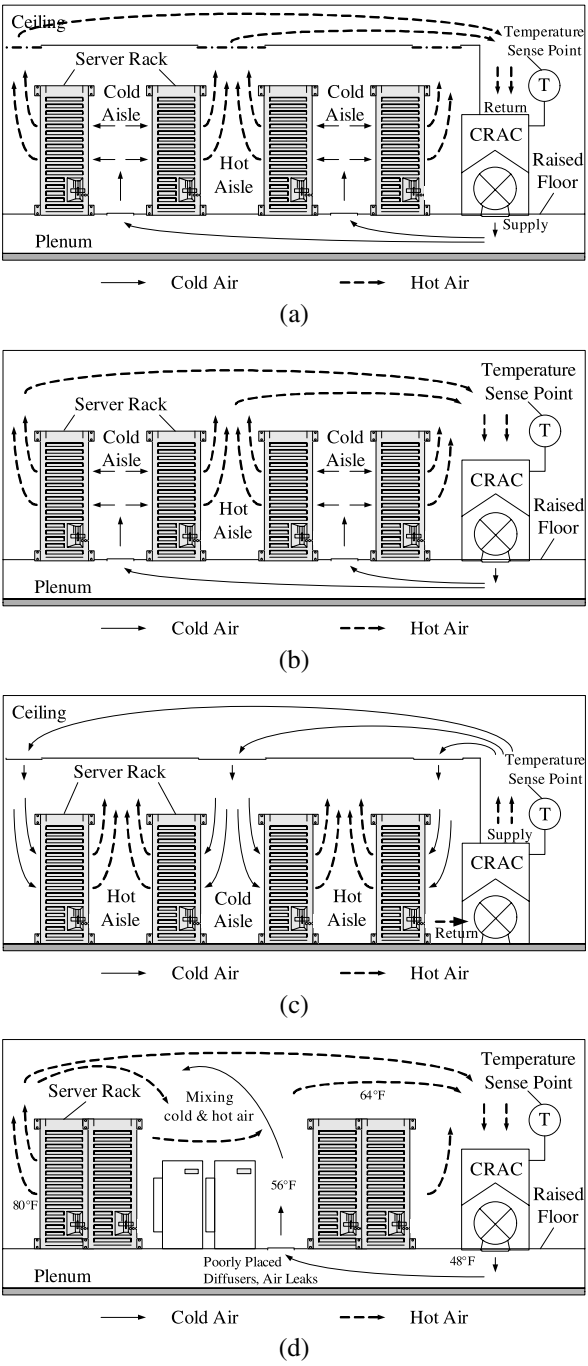

Fig. 1 Current cooling design schemes: (**a**) underfloor supply and overhead return, (**b**) underfloor supply and horizontal return, (**c**) overhead supply with horizontal return, and (d) poor airflow condition

racks with the objective of increasing the overall efficiency of the air delivery and collection from each rack in the datacenter. Different thermal efficiencies may be achieved with alternate configurations, as illustrated in Figs. 1(b) and 1(c).

Any configuration for the CRAC unit can be applied, but only certain combinations are feasible due to mechanical constraints of the CRAC units (e.g., to not introduce an excessive amount of duct work). With these constraints, achieving thermal (energy) efficiency is complicated and there is no single optimal solution. In [17], the authors made an attempt to compare cooling efficiencies among four airflow distribution systems in high heat density room: underfloor supply/overhead return, underfloor supply/horizontal return, overhead supply/underfloor return, and overhead supply/horizontal return. In [8, 9], the authors expand on the concepts proposed in [17]. Datacenter design guidelines from PG&E [19]—mainly based on [2]—presents a poorly designed datacenter room (Fig. 1(d)) cooled by a raised floor system, which has often trouble maintaining an appropriate room temperature. In [7], the authors benchmark 22 datacenters according to their energy usage and conclude that energy benchmarking using a specific metric—ranging from the energy used for IT equipment to the energy used for Heating, Ventilation, and Air Conditioning (HVAC)—is extremely helpful to understand why some datacenters perform better than others.

Another important aspect of mechanical design for cooling system involves airflow distribution. In [12, 22, 23, 25], the authors provide some insights into the airflow distribution from perforated tiles and raised floor design in datacenters. In [21, 24], the authors thermally profiled a datacenter in space and time, and analyzed trends and correlations among collected measurements. Basic mathematical modeling and parameters for modeling datacenter are proposed in [27]. In [25], the authors comment on the challenges associated with thermal management in datacenters. In [20], the authors review the existing literature on datacenter thermal modeling work. However, due to the complexity of the thermodynamics [3, 18], the research done in formulating models suitable to describe the complex phenomena of heat propagation and air distribution in datacenters has been limited.

On the other hand, software-based approaches focus on minimizing the cooling cost by distributing or migrating jobs. In [10], the authors emulate the thermal behavior of server systems to manage datacenter cooling. Using the emulator, a system named Freon monitors temperature changes and, if the temperature of a machine crosses a threshold (defined thermal emergency), Freon redistributes the jobs; however, cooling cost is not considered in this study. In [15], the authors introduce the concept of power budget, which is the product of power and temperature. Higher power budget means that a machine has more capacity to accept a job in terms of temperature and power. The authors propose two scheduling algorithms based on power budget to fairly and efficiently distribute the workload. In [16], the authors propose energy savings by temporally spreading the workload and assigning it to energy-efficient computing equipment. However, these works are not coupled to a physical datacenter model and only consider scenarios that are reactive in nature. In [29], the authors propose a cooperative power-aware game theoretic solution for job scheduling in grids to minimize the energy consumption while maintaining a specific Quality of Service (QoS) level. They highlight the fact that it is not enough to minimize the total energy of grid but there is the need to minimize energy locally at

different providers in the grid. The proposed solution simultaneously minimizes the energy used by all providers so to be fair to all users. The energy usage is kept to minimum level while maintaining the desired QoS level. The authors claim that the proposed solution is robust against prediction inaccuracies. In [13], the authors study the problem of task allocation onto a computational grid and aim at simultaneously minimizing the energy consumption and the makespan (time difference between the start and finish of a sequence of jobs), subject to the constraints of deadlines and tasks' architectural requirements. The solution is proposed from cooperative game theory based on the concept of Nash Bargaining Solution (NBS). The frameworks proposed in [13, 29] do not take the datacenter design or airflow characteristics into consideration and are, hence, not suitable for optimizing the cooling system performance.

In current datacenter thermal management, the mechanical- and software-based approaches are usually independent on each other. However, there exists a strong correlation between the two; for this reason, there is the need to combine the two approaches to obtain an optimal cooling solution. The proactive approach proposed in this work can adapt itself to any type of mechanical design and it also considers the distribution and type of workload running. It combines the mechanical aspect of a datacenter with the software-based scheduling approach to optimize the performance of the cooling system.

3 Problem formulation

In this section, we formulate the mathematical model for heat transfer in datacenters that our solution is based on. The heat transfer model is divided into three parts as follows: (i) overall heat circulation model, (ii) heat generation model, and (iii) heat extraction model. We describe heat generation, extraction, and circulation based on the fundamental thermodynamic principles in physics. It is assumed that the datacenter is built on a hot-and-cold aisle based raised plenum design, supplying cold air from the raised plenum and returning hot air to the ceiling.

3.1 Air circulation design

The proposed datacenter model is designed in a 3-dimensional space as shown in Fig. 2. In Fig. 2, the columns represent aisles, the rows represent the distance from the CRAC, and the height represents the enclosure number from the bottom. These are referred to as "i," "j," and "e," respectively. The latest rack design includes 3–4 enclosures, each containing 10–20 integrated vertical blades with an independent cooling module to cool all the blades in it. We assume that every odd numbered aisle is a cold aisle and every even numbered aisle is hot aisle. For the sake of clarity, the notation used in the model are summarized in Table 1.

The proposed model is based on heat imbalance equations. Heat is mainly generated by the processor and subsystems, i.e., memory and storage devices, I/O subsystem, network interface card, etc., is extracted by the fans in the enclosure and the fan in the CRAC unit. We assume that there is no external heat source and the room is

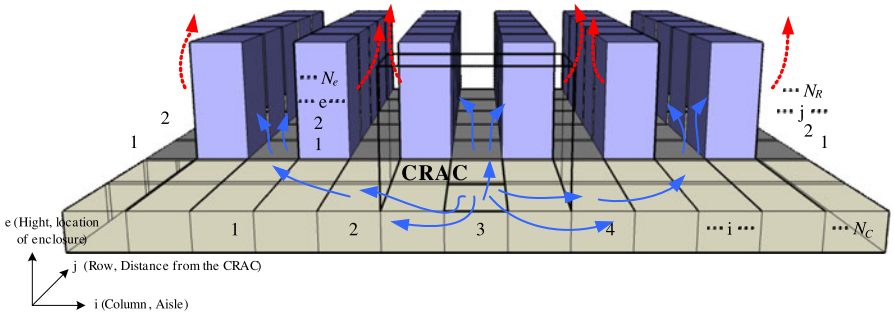

Fig. 2 Layout of datacenter and location of a blade

thermally insulated. The notations used to describe the datacenter thermal flow model is similar to the ones used in [27].

In order to calculate the heat imbalance and flux, we assume that our datacenter is extended in a 2-dimensional space as shown in Fig. 3(a) and explained following airflow from the CRAC ("①" in Fig. 3(a)) and returning to CRAC ("⑤–⑥" in Fig. 3(a)). First, a cold air stream from the CRAC unit is pumped through the plenum at flow rate $m_{\text{crac}}^{\text{out}}$ and temperature T_{crac} in "①"; this flow is evenly divided and exhausted through the plenum and perforated tiles "②" in the cold aisle ideally. However, in real case, perforated tiles can be modeled as a lumped resistance using the relationship $\Delta P = \beta \cdot (m_{\text{crac}}^{\text{out}}/\rho)^2$, where the coefficient β can be found in standard flow resistance handbooks (e.g., [26]). Experimental values of β have been proposed in [21, 24]. Using equation for *dynamic pressure* $\Delta P = \rho \cdot v^2/2$, where v is the fluid velocity, $m_{\text{tile}}^{\text{in}}$ and $m_{\text{tile}}^{\text{out}}$ can be written as follows:

$$m_{\text{tile}}^{\text{in}} = m_{\text{tile}}^{\text{out}} = A_{\text{tile}} \cdot \sqrt{\beta \cdot \frac{2 \cdot (m_{\text{crac}}^{\text{out}})^2}{\rho}}. \tag{1}$$

Between "②" and "③," inlet airflow rate of an enclosure is proportional to inlet airflow of a rack. Ideally, if there is no air circulation, leakage or Bernoulli effect,[1] and the fans on every enclosure have the same speed then, m_e^{in} is m_r^{in}/N_E and m_r^{in} is half of $m_{\text{tile}}^{\text{out}}$ because the air flowing from a tile splits into two racks. The relations of those parameters are

$$m_r^{\text{in}} = \sum_{e=1}^{N_E} m_e^{\text{in}} = N_E \cdot m_e^{\text{in}} = \frac{m_{\text{tile}}^{\text{out}}}{2}. \tag{2}$$

In ideal case, the inlet mass airflow rate in through "③" for all the racks are the same as

$$m_{\text{in}}^r = m_{\text{in}}^l, \quad \forall_{r,l}. \tag{3}$$

[1] In fluid dynamics, the Bernoulli's principle states that an increase in the speed of the fluid occurs simultaneously with a decrease in pressure or a decrease in the fluid's potential energy.

Table 1 Notations

Nomenclature	
ΔI	Heat imbalance [J]
C_p	Specific heat of air at constant pressure [J/kg K]
m	Mass airflow rate [kg/s]
M	Mass of air [kg]
T	Temperature [K]
ρ	Density of air [kg/m^3]
A	Area of the component [m^2]
N_C	Total number of columns/aisles
N_R	Total number of rows
N_A	Total number of Rack
N_E	Total number of enclosures in a rack
Subscript	
e	Enclosure
r	Rack
i	Column
j	Row
crac	CRAC unit
tile	Perforated vent tile
room	Room where the equipment is placed in a datacenter
Superscript	
in	Inlet
out	Outlet

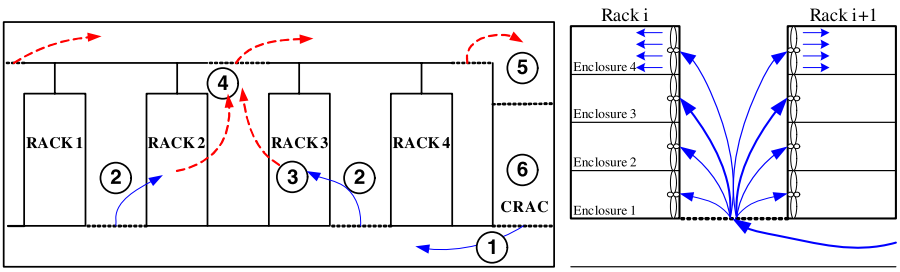

Fig. 3 (**a**) 6 different places to consider mathematical model; (**b**) Rack level airflow

As these ideal assumptions are often not valid in realistic scenarios, to improve the accuracy of the model we need realtime measurements collected via a sensing infrastructure, as discussed in Sect. 6. Figure 3(b) shows the flow from the tile to the

rack and (2) shows the relation between $m^{\text{out}}_{\text{tile}}$ and m^{in}_{r}, where "r" denotes the racks between aisle i to $i+1$ (column) and j (row) cell. Cooling fans in each enclosure suck the air streams in at flow rate m^{in}_{e}, cool the heated components down in "③," and the air stream in the enclosure flows to the back of the racks with flow rate m^{out}_{r} in "④." We formulate *heat-imbalance model* in a datacenter as follows, which explains the heat exchange in "③" as

$$\Delta I^{e}_{ij} = \int_{t_1}^{t_2} \left(h^{e}_{ij} - q^{e}_{ij}\right) dt = M^{e} \cdot C_p \cdot \Delta T^{e}_{[t_2,t_1]}, \tag{4}$$

where

- ΔI^{e}_{ij} denotes the heat imbalance of the enclosure e in the cell (i, j), which is between i and $i+1$, and j, during the time between t_1 and t_2;
- h^{e}_{ij} is the rate of heat generation from the enclosure e in the cell (i, j) [J/s];
- q^{e}_{ij} is the rate of heat extraction from the enclosure e in the cell (i, j) [J/s];
- If ΔI^{e}_{ij} is positive (i.e., $h^{e}_{ij} > q^{e}_{ij}$), the temperature in the enclosure e increases (hence, $\Delta T^{e} > 0$);
- If ΔI^{e}_{ij} is negative (i.e., $h^{e}_{ij} < q^{e}_{ij}$), the temperature in the enclosure e decreases (hence, $\Delta T^{e} < 0$).

Equation (4) shows the difference between heat generated and heat extracted in an enclosure. If the heat difference is positive, the enclosure temperature will increase; if the imbalanced heat of the entire datacenter is set up as a function of blades of the enclosures, and enclosures of the racks, then we have

$$\Delta I = \sum_{i=1}^{N_C} \sum_{j=1}^{N_R} \sum_{e=1}^{N_E} \Delta I^{e}_{ij} = M_{\text{room}} \cdot C_p \cdot \Delta T^{\text{room}}_{[t_2,t_1]}. \tag{5}$$

If the heat difference is positive, the average temperature of the datacenter will increase.

From the experiments on our test server, which has the following configuration, eight 8 core Intel Nehalem processors, 138 GB of RAM and 500 GB of storage, it was observed that the dominant power utilizing subsystems were the CPU, I/O subsystem, memory and storage subsystem, and the Network Interface Card (NIC). Out of the total power utilized, a certain percentage was dissipated as heat. The percentage of power dissipated as heat by the subsystems is denoted by α^{sub} [%] as detailed in Sect. 4.1. The rate of heat generation by a subsystem h^{e}_{ij} [J] is given by

$$h^{e}_{ij} = P^{e,\text{cpu}}_{ij} \cdot \alpha^{\text{cpu}} + P^{e,\text{IO}}_{ij} \cdot \alpha^{\text{IO}} + P^{e,\text{mem,stg}}_{ij} \cdot \alpha^{\text{mem,stg}} + P^{e,\text{NIC}}_{ij} \cdot \alpha^{\text{NIC}}, \tag{6}$$

where P^{cpu}, P^{IO}, $P^{\text{mem,stg}}$, P^{NIC} is the power utilized by CPU, I/O subsystem, memory and storage devices, and the NIC, respectively, and α^{cpu}, α^{IO}, $\alpha^{\text{mem,stg}}$, α^{NIC} are the respective percentage power dissipation factors. Using (29), the total rate of heat generation "H" can be calculated as,

$$H = \sum_{i=1}^{N_C} \sum_{j=1}^{N_R} \sum_{e=1}^{N_E} h^{e}_{ij}. \tag{7}$$

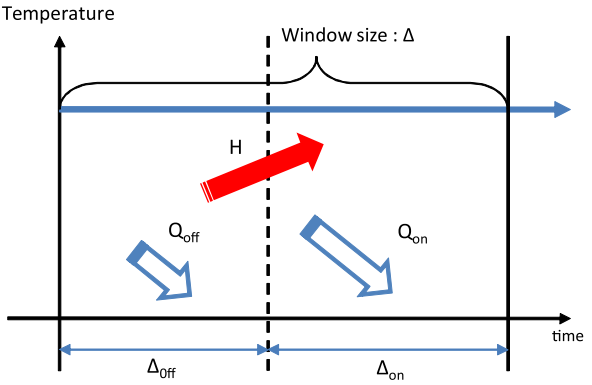

Fig. 4 Q_{on} is the heat extraction when compressor is on, Q_{off} is the heat extraction when compressor is off which are proportional to $T_{\text{out}} - T^{\text{crac}}$ and T_{out} - T^{room}, respectively, while H is the heat generation.

On the other hand, the heat is extracted by the inlet-air and flows out with the outlet-air, which can be calculated as

$$q_{ij}^{e} = m_{ij,\text{in}}^{e} \cdot C_p \cdot \left(T_{ij,\text{out}}^{e} - T_{ij,\text{in}}^{e}\right). \tag{8}$$

By measuring inlet and outlet temperature and airflows, we can calculate how much heat is extracted from the enclosure e located in i (row) and j (column), i.e., in cell (i, j). The total rate of heat extracted can be computed as,

$$Q = \sum_{i=1}^{N_C} \sum_{j=1}^{N_R} \sum_{e=1}^{N_E} q_{ij}^{e}. \tag{9}$$

Temperature of the inlet airflow $T_{ij,\text{in}}^{e}$ varies depending on whether the air compressor is "on" or "off" in (8). In Fig. 4, if the air compressor is on (Q_{off}), the air from the server room with the temperature (T^{room}) can extract the heat because only the fan is working, but if the air compressor is on (Q_{on}), then the air with the temperature from the CRAC unit (T^{crac}) can extract the heat because both the compressor and fan are working. The heat generation "H" is independent on whether air compressor is working or not because it is generated based on the workload and it's distribution. In equilibrium state, 'T^{crac}' is lower than 'T^{room}' and 'T^{room}' is lower than 'T_{out}' ($T^{\text{crac}} < T^{\text{room}} < T_{\text{out}}$).

Heated air streams are collected through the ceiling in "⑤" and returned to the CRAC unit. Finally, the circulated air enters the CRAC unit and is compressed and cooled again in "⑥." In an ideal case, it holds,

$$m_{\text{out}}^{\text{crac}} = \sum_{r=1}^{N_A} m_{\text{in}}^{r} = \sum_{r=1}^{N_A} m_{\text{out}}^{r} = m_{\text{in}}^{\text{crac}}. \tag{10}$$

The air can be mixed for a variety of reasons, some of which are discussed in Sect. 2. The phenomenon and the effect of mixing cold air and hot air streams are discussed

in [27]. Supply using Heat Index (SHI) is denoted as follows:

$$\mathrm{SHI} = \frac{\sum_{i=1}^{N_C} \sum_{j=1}^{N_R} (T^r_{\mathrm{in},ij} - T^{\mathrm{crac}}_{\mathrm{out}})}{\sum_{i=1}^{N_C} \sum_{j=1}^{N_R} (T^r_{\mathrm{out},ij} - T^{\mathrm{crac}}_{\mathrm{out}})}. \tag{11}$$

Because the air from the CRAC unit $T^{\mathrm{crac}}_{\mathrm{out}}$ is not at the same temperature as the rack T^r_{ij}, we can assume that there is a recirculation and mixing of hot–cold air in the datacenter. We can apply the simple SHI configuration for our model to explain air recirculation. However, this model using inlet and outlet airflow can be applied only if we know the inlet and outlet temperature and airflows at each blade. This requires a sensing infrastructure that measures airflows and temperature to quantify the amount of heat extracted.

3.2 Air cooling system design

Cooling is the process where heat is transferred from a lower to a higher temperature level by doing work on a system in order to extract the heat. Most datacenters use chilled-water air conditioning system. With chilled water air conditioning, the refrigeration equipment (compressor, condenser, evaporator, etc.) does not directly cool the air; rather, it uses chilled water to cool the air, where chilled water is pumped to the cooling coils and a fan draws the air through the chilled water pipe to the coils, thus cooling the air. With chilled water air conditioning, the compressor is usually mounted on a rack or a frame, within a few feet from the evaporator that cools the chilled water. The efficiency of this cycle can be determined by several factors such as airflow and chilled water temperature, and can be quantified by the Coefficient Of Performance (COP). The COP is the ratio of total heat removed 'Q' from low-temperature level and the energy input used (W) as

$$\mathrm{COP}(T) = \frac{Q}{W}. \tag{12}$$

Since the COP is inversely proportional to the W, a higher COP means that more heat "Q" can be removed by doing less work (W), as given in (12). The COP can be calculated in each cooling cycle and through (12), we can calculate how much work is needed to extract a certain amount of heat. For example, a cooling cycle with COP of 2 will consume 30 kWh to extract 60 kWh of heat, while a cooling cycle of COP of 3 will consume 20 kwh to remove the same amount of heat. Figure 5 shows COP values for different CRAC units. As the CRAC supply temperature increases, the COP also increases (in compliance with the second principle of thermodynamics). Consequently, our cooling cost "E_{Total}" can be calculated as

$$E_{\mathrm{Total}} = E_{\mathrm{Compressor}} + E_{\mathrm{fan}}, \tag{13}$$

which is the total amount of energy needed to power the air compressor, "$E_{\mathrm{Compressor}}$," plus energy needed to run the CRAC fan, "E_{fan}." If the duty cycle η of the compressor is represented as

$$\eta = \frac{\Delta t_{\mathrm{on}}}{\Delta t_{\mathrm{on}} + \Delta t_{\mathrm{off}}}, \tag{14}$$

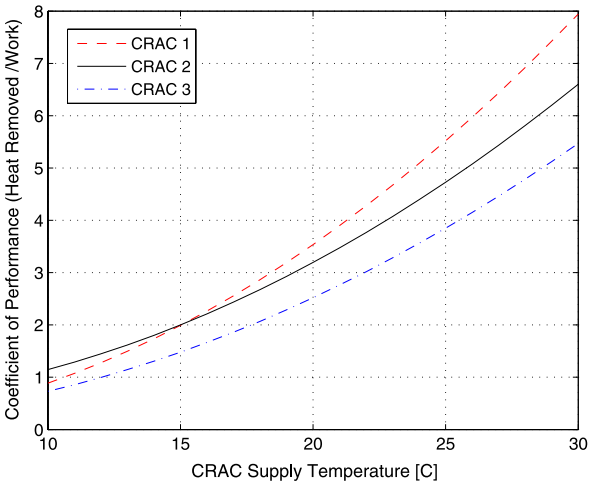

Fig. 5 Coefficient of Performance (COP) curve for the different chilled-water CRAC units. COP can vary for different CRAC unit depending on the type of fan and compressor used. CRAC 2 shows COP of cooling system at the HP Labs Utility Datacenter [15]. This COP is also used in our simulations

where the window size of the compressor cycle is represented as "$\Delta = \Delta t_{\text{off}} + \Delta t_{\text{on}}$," then the work done to extract the amount of heat for a compressor can be calculated as

$$E_{\text{Compressor}} = P_{\text{AC}} \cdot \Delta t_{\text{on}} = \frac{Q}{\text{COP}(T)}. \tag{15}$$

The "*affinity law*," which describes how the performance of a centrifugal pump is affected by a change in speed or impeller diameter is given by

$$P = P_{\text{ref}} \cdot \frac{\omega^3}{\omega_{\text{ref}}^3}, \tag{16}$$

where ω is the shaft rotation speed (fan speed) and P is the shaft power correspond to the ω. Note that the shaft power is proportional to the cube of rotation speed of the fan shaft. "E_{fan}" can be calculated using this law. Specifically, once we know the reference-point power P_{ref} of the fan and its rotational speed ω_{ref}, which both vary depending on the manufacturer and type of fan, the power required to increase the fan speed to ω in $[t_1, t_2]$ can be computed as

$$E_{\text{fan}} = \int_{t1}^{t2} P_{\text{ref}} \cdot \frac{\omega^3}{\omega_{\text{ref}}^3}\, dt. \tag{17}$$

Also, the mass of airflow rate injected from the CRAC, "m_{crac}," can be computed based on fan speed ω as

$$m_{\text{crac}} = \rho(T) \cdot K_\omega \cdot \omega, \tag{18}$$

where $\rho(T)$ denotes density of the air in certain temperature and K_ω denotes coefficient of amount of air through the fan, when the speed of the fan is ω.

Since there is only one fan with airflow rate "m_{crac}" is used in this model, we cannot extract the localized heat generated in each aisle or each rack. However, if we

use multiple fans, then the total mass airflow rate in each corridor is given by

$$m_{\text{crac}} = m_{\text{fan}_1} + m_{\text{fan}_2} + m_{\text{fan}_3} \ldots, \tag{19}$$

where $m_{\text{fan}_{\text{num}}}$ denotes the mass airflow rate of each fan. If we install multiple fans, one for each corridor or each rack, then we can have more control over the net airflow rate. However, there is an additional cost for installation and operation of these fans. Power usage by a fan ranges from hundreds to thousands Watt and the power usage by the air compressor is hundreds of kilowatt. However, the energy savings obtained by just increasing fan speed and not increasing the compressor cycle to extract localized heat offsets the additional cost. In this way, we can selectively extract localized heat without much extra cost.

3.3 Problem formulation to minimize cooling energy

We formulate the problem to optimize fan speed (ω^*) and duty cycle of the air compressor (η^*) to minimize the energy consumed by the compressor and fan as follows.

$$\begin{aligned}
\text{Given (offline):} \quad & T^{\text{crac}}, T_{\text{set}}, T_{\text{init}}, P_{\text{AC}}, \omega_{\text{ref}}, P_{\text{ref}}, M_{\text{room}}, \Delta, C_p, \rho(), \\
& \text{COP}(), K_\omega, N_C, N_R, N_E \\
\text{Given (online):} \quad & h^e_{ij}, T_{\text{out}}, T^{\text{room}} \\
\text{Find:} \quad & \omega^*, \ \eta^* \\
\text{Minimize:} \quad & E_{\text{Total}} = E_{\text{Compressor}} + E_{\text{fan}} = P_{\text{AC}} \cdot \eta \cdot \Delta + P_{\text{ref}} \cdot \frac{\omega^3}{\omega^3_{\text{ref}}} \\
\text{Subject To:} \quad &
\end{aligned}$$

$$\Delta I = \int_{t_0}^{t_0+\Delta} H \, dt - Q_{\text{on}} - Q_{\text{off}} = M_{\text{room}} \cdot C_p \cdot \left(T_{\text{set}} - T_{\text{init}}\right); \tag{20}$$

$$H = \sum_{i=1}^{N_C} \sum_{j=1}^{N_R} \sum_{e=1}^{N_E} h^e_{ij}; \tag{21}$$

$$Q_{\text{on}} = \rho \cdot K_\omega \cdot C_p \cdot \left(T_{\text{out}} - T^{\text{crac}}\right) \cdot \eta \cdot \Delta; \tag{22}$$

$$Q_{\text{off}} = \rho \cdot K_\omega \cdot C_p \cdot (T_{\text{out}} - T^{\text{room}}) \cdot \eta \cdot (1 - \Delta). \tag{23}$$

Constraint (20) forces the heat imbalance "I" to be equal to the amount of heat to adjust the temperature to the set point from the initial temperature, so that the server room temperature remains in equilibrium with the set point. By using this constraint and on-line and off-line parameters obtained from the datacenter, the fan speed ω and compressor cycle η can be optimized. If the amount of heat generated is the same

as the amount of heat extracted, then there is no heat unbalance and the temperature stays in the equilibrium point. Equation (21) shows the heat generated from each component of the server blade. Equation (22) shows heat extraction when the compressor is turned on, and (23) shows heat extraction when the compressor is turned off.

4 Proposed solution

In this section, we propose our proactive approach model and later we describe the reactive approach. The proactive approach keeps the return temperature or internal blade-temperature in a safe operating range by jointly optimizing the duty cycle of air compressor and the CRAC fan speed before the rack/enclosures are heated using knowledge of the workload. The reactive approach is based on a feedback mechanism, in which the external temperature is adjusted back to a safe operating range when the heat generated by the rack/enclosures causes heat imbalance leading the external temperature to rise above a certain safe operating threshold temperature.

4.1 Proactive approach

The proactive approach solves the problem at grass-root level. It is based on temperature change due to the heat imbalance between the heat generated and heat extracted. This proactive approach is *quantitative* in nature as it measures/estimates the heat imbalance. Conversely, compared to a proactive approach, a reactive approach is only *qualitative* in nature as it reacts to the changes in temperature. Using an analogy with kinematics in physics, we can say that *a proactive approach is analogous to determining the final position of an object by measuring its velocity instead of the position itself.*

The effectiveness of a proactive approach is based on having comprehensive knowledge about the behavior of the workload running and its utilization of the subsystems in a blade. Knowing the subsystem usage pattern of the workload that is expected to run in advance and knowing the specifications of the subsystems by the application/workload layer (Fig. 6), the heat dissipated by each of the subsystems can be estimated. The subsystem usage pattern is obtained by observing the subsystem usage behavior of the selected workload over a certain period of time. The historical data obtained is used to generate workload patterns as shown in Fig. 7. These patterns can be used to estimate the power utilized and, in turn, the heat generated from the subsystems. Estimating the heat generated by subsystems and knowing the heat extracted by the cooling system, i.e., the heat imbalance, the temperature rise at a blade can be estimated. Based on this predicted temperature rise, the external temperature can be adjusted accordingly in order to maintain the internal temperature under safe operating threshold by the Environment/Physical Layer, as in the schematic in Fig. 6.

Based upon the subsystem usage pattern of the workload provided by the Application/Workload Layer, we propose two solutions; (i) proactive approach using single fan, and (ii) proactive approach using multiple fans. Single fan approach uses one fan for the CRAC unit; however, single fan approach could be too global to adjust

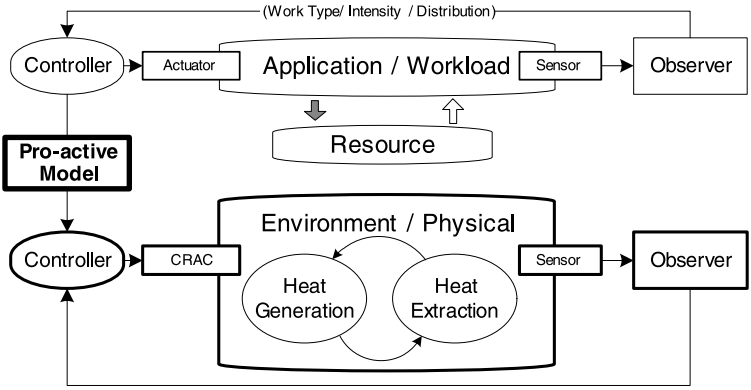

Fig. 6 Proactive approach

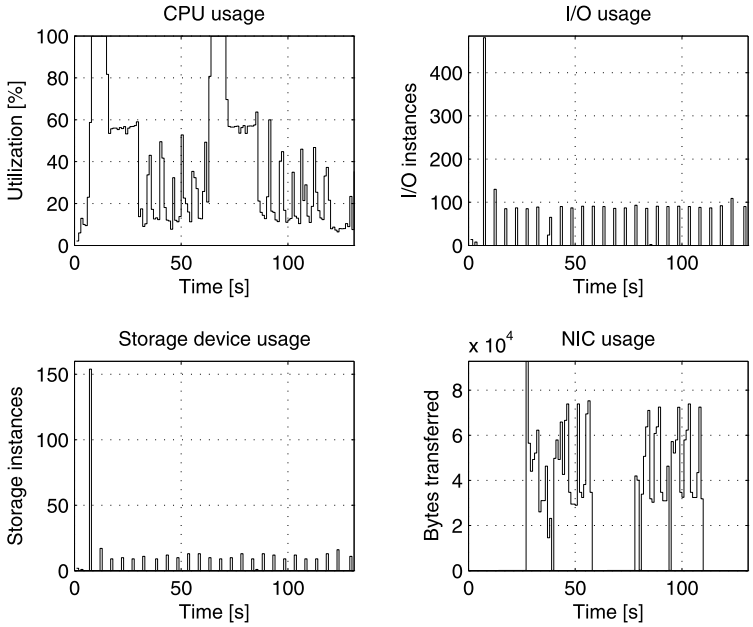

Fig. 7 FFTW benchmark subsystem usage profile showing processor, I/O subsystem, memory subsystem, and the NIC usage

local temperatures since only few estimations of overheating can trigger the entire CRAC unit to operate. This inefficiency may be reduced by the use of multiple fans assigned one for each aisle. For example, if the number of jobs assigned in a certain aisle is more than other aisles, we can optimize the fan speed based on the workload type and pattern. Since the energy needed for adjusting fan speed is much lower than the energy needed for adjusting the duty cycle of air compressor, we can reduce the energy consumption.

We obtain the historic subsystem usage data for some standard HPC benchmark workloads like, FFTW, HPL, NAS-benchmarks, from our test server to observe their subsystem usage pattern. The data in Fig. 7 shows the subsystem usage pattern of the *FFT* workload for CPU, I/O subsystem, storage device, and NIC utilization with respect to time in four subplots. Each subplot represents the magnitude of usage of a subsystem with respect to time. CPU usage is represented as percentage usage per second, I/O and storage usage is represented in the units of input/output instances or read/write instances per unit time, and the NIC usage is represented in units of data bytes exchanged per unit time using the TCP protocol. The subsystem usage pattern of the workload provides us with the information about the time instances at which each subsystem is utilized and when it is idle. From the pattern in Fig. 7, we can estimate the power utilized by the subsystems and, in turn, the heat generated. The subsystem usage patterns are the back bone for estimating the heat generated at the blade level.

Combining the power dissipated as heat at the CPU and other subsystems, the total power dissipation of the blade is estimated. All the subsystems are composed of semiconductor devices, hence we calculate the leakage power dissipated as heat from the formulas given in [5]. This is an approximate estimate as modeling the exact amount of heat dissipated is complicated and not absolutely necessary. In the model, we are interested in estimating the maximum heat that can be generated at any of the subsystems at any time instant based on the workload pattern. Hence, we assume it is safe to use the leakage power formulas presented in [5].

The leakage power P_{leakage} for a semiconductor chip as given in [5] is

$$P_{\text{leakage}} = W_{\text{avg}} \cdot I_{\text{leak}} \cdot N_{\text{trans}} \cdot V_{\text{dd}}, \tag{24}$$

where W_{avg} is the average size of the transistor at the input gate, I_{leak} is the leakage current per unit width, N_{trans} is the total number of transistors in "on" state, V_{dd} is the input voltage. In (24), N_{trans} is dependent on the device usage, which is obtained from the workload pattern, all the remaining parameters are obtained from the technical specifications of the motherboard and individual IC datasheets. Hence, (24) provides us with the direct relation between the subsystem utilization and heat dissipation. Also, some of the subsystems like the CPU have multiple sleep states also known as C-states and the value of N_{trans} varies depending on the power state in use, while other subsystems may have only two states, i.e., "on" and "off."

The test server is configured to have the following specific configuration: (1) Advanced Configuration and Power Interface (ACPI) is enabled, and (2) processor dynamic frequency scaling is enabled. Since ACPI is enabled, the CPU can transition to multiple C-states, with C_0 being the most power utilizing state or an active state, and C_n being the least power utilizing state or a deep sleep state [1]. Other subsystems, i.e., I/O subsystem, memory and storage devices, and the NIC, do not have any operating system based power management enabled, and hence have only two states with D_0 being the "on" state (or most power utilizing state) and D_n being the "sleep" state (or least power utilizing state). With dynamic frequency scaling enabled, the CPU can transition to various predetermined frequency levels also known as P-states. Because the CPU has multiple sleep and power states, we calculate its power separately from other subsystems.

Power utilization of the processor in C_0 is a function of P-states or the frequency at which the processor is running. The power utilization of the processor is calculated as

$$P^{\text{cpu}} = P_{C_0}^{\text{cpu}} + \cdots + P_{C_n}^{\text{cpu}}, \tag{25}$$

where $P_{C_0}^{\text{cpu}}$ is given as

$$P_{C_0}^{\text{cpu}} = \sum_{j=0}^{k} P_{P_j}, \tag{26}$$

where P_{P_j} [W] is the power utilized in the P-state j assuming processor has k P-states. P-states are not relevant for sleep states (C-states other than C_0) as processor is inactive in them. Hence, the power utilized in C-states other than C_0 is given as $P_{C_n}^{\text{cpu}}$ [W], where "n" is the C-state depth. The power utilized by other subsystems, i.e., I/O subsystem, storage devices, and the NIC, is given by

$$P^{\text{sub}} = P_{S_0}^{\text{sub}} + P_{S_{\text{idle}}}^{\text{sub}}, \tag{27}$$

where P^{sub} [W] is the total power utilized by the subsystem and $P_{S_0}^{\text{sub}}$ [W] is the power utilized when the subsystem is in use or "on" sate and $P_{S_{\text{idle}}}^{\text{sub}}$ [W] is the power consumed when the subsystem is idle or in "sleep" state.

The percentage of power dissipated as heat $\alpha^{\text{cpu,sub}}$ [%] is given by,

$$\alpha^{\text{cpu,sub}} = \frac{P_{\text{leakage}}^{\text{cpu,sub}}}{P^{\text{cpu,sub}}} \times 100, \tag{28}$$

where $P^{\text{cpu,sub}}$ is the total power utilized by the CPU and subsystems, and $P_{\text{leakage}}^{\text{cpu,sub}}$ is the leakage power for the CPU and subsystem calculated from (24).

Total Heat generated h_{ij}^{e} [J] by the processor and subsystems in blade, over the time $t^{\text{cpu,sub}}$ is given by

$$h = P^{\text{cpu,sub}} \cdot \alpha_d^{\text{cpu,sub}} \cdot t^{\text{cpu,sub}}. \tag{29}$$

From (29), we observe that h_{ij}^{e} is directly proportional to the product of power utilized by the processor and subsystems $P^{\text{cpu,sub}}$ and the time $t^{\text{cpu,sub}}$ when the processor and subsystems are "on."

4.2 Reactive approach

The reactive approach is based on measuring the change in temperature and accordingly adjusting the duty cycle and fan speed of the CRAC unit in Fig. 8. Reactive approach based models can be implemented in two ways depending on where the temperature change is measured: (i) by measuring the change in return temperature, and (ii) by measuring the change in internal temperature at each blade. There are advantages and disadvantages for both approaches. Since the former approach uses only the return air temperature, it is simple to adjust the controller. However, it could

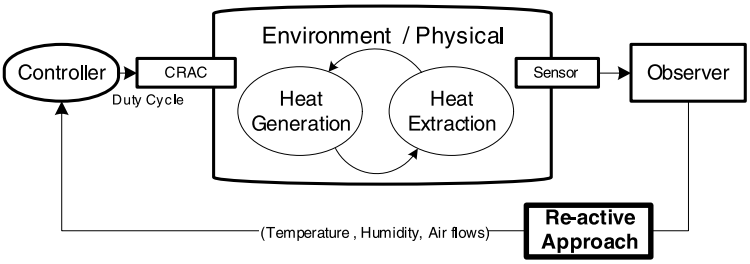

Fig. 8 Reactive approach

have a substantial delay depending on how fast the room air is circulated and how big the machine room size is.

The advantages of the reactive approach is that we are directly observing the temperature at the place where we want to control the temperature. If the temperature of any blade rises up to the critical point that can damage the hardware components, air conditioning system reacts to cool down the system. Controller can react faster in this approach than the former one because the source of the problem is close to the air conditioning system, but it still has delay and cooling can be more expensive as only a few overheated machines can trigger an entire cooling system. Moreover, it increases the complexity of the control and communication mechanism and this approach measures the change in temperature but does not measure the quantitative heat generated. Hence, this approach is inefficient compared to proactive approach.

The reactive approach takes corrective actions after the temperature has crossed a threshold temperature $TH_{\text{high,low}}$. One possible action is to increase the fan speed, which leads to increase the flow of the air and extract more heat from the blades. As the heat extracted depends on the inlet temperature, if the temperature of the room is not low enough to cool down the machine this action is not affective. Another possible action is to increase the compressor duty cycle so to lower the temperature of the air. Within the reactive approach, however, a fine balance between these two actions as well as the optimal tuning of the parameters controlling them are not possible as quantitative relations between causes and effects are missing. For this reason, we proposed the proactive approach, whose performance is provided in the following section.

5 Performance evaluation

In this section, we analyze the performance of the models developed in this work. The simulations are built using MATLAB®. The nonlinear optimization problem in Sect. 3.3 is solved using the *fmincon* solver in the Optimization Toolbox of MATLAB®. In Sect. 5.2, we analyze the proactive approach based mathematical model using the chosen benchmarks and compare the results with that of the reactive approach. The proactive approach using multiple fans is also simulated. In Sect. 5.3, we compare the energy consumption and the risk of overheating for both reactive and proactive approaches.

Table 2 List of benchmarks used

Benchmark Name	Benchmarks Type
FFTW	Computing discrete Fourier Transforms (Intensive)
NAS-SP	Benchmark from the NASA Parallel Benchmarks (NPB) family (Benchmark)
HPL Linpack	Solves a linear system on distributed-memory computer (Compute)

Table 3 Input variables for the simulation

Input	
Maximum airflow from the CRAC	18000 m^3/h
Supply-temperature from the CRAC	17°C
Number of blades	1260
Number of enclosures	90
Number of racks	30
Number of perforated tiles	15
Size of a perforated tile (A_{tile})	1 m^2
Mass of the air in the datacenter	220.5 kg
Maximum fan flow	5 m^3/s
Set point	22°C
threshold for the risk condition	70°C
Heat generation model	HPL Linpack, FFTW, NAS-SP
Scheduling algorithm	Random load balancing, Sequential load balancing

Table 2 shows the benchmarks used to validate the model. The chosen workload benchmarks are compute intensive and generate substantial amount of heat. Most of the workloads that run in high performance clusters and the datacenters are compute intensive, and are comparable to the chosen workload benchmarks. Using the chosen benchmarks and input variables in Table 3, we simulate the thermal behavior of a datacenter and the "heat imbalance (Q)." We used two types of scheduling algorithms for simulations: (i) random load balancing and (ii) sequential load balancing. Random load balancing selects random blades and assigns workloads, whereas sequential load balancing selects sequential blades that are closely located and assigns workloads accordingly.

The chosen benchmark workloads were run and profiled on our test server. Their processor, I/O subsystem, memory subsystem, and NIC usage with respect to time were measured using custom scripts. We obtained the time for which each subsystem was "on" from the profiling data, and from the data sheet we know the power utilized by the subsystems and also the power dissipated. With this information, we derived the heat dissipated per unit time based on this profile data.

5.1 Reactive approach

In reactive approach, either the "duty cycle of a compressor (η)" or "fan speed of the CRAC (ω)" can be changed based upon the user specifications. We choose the

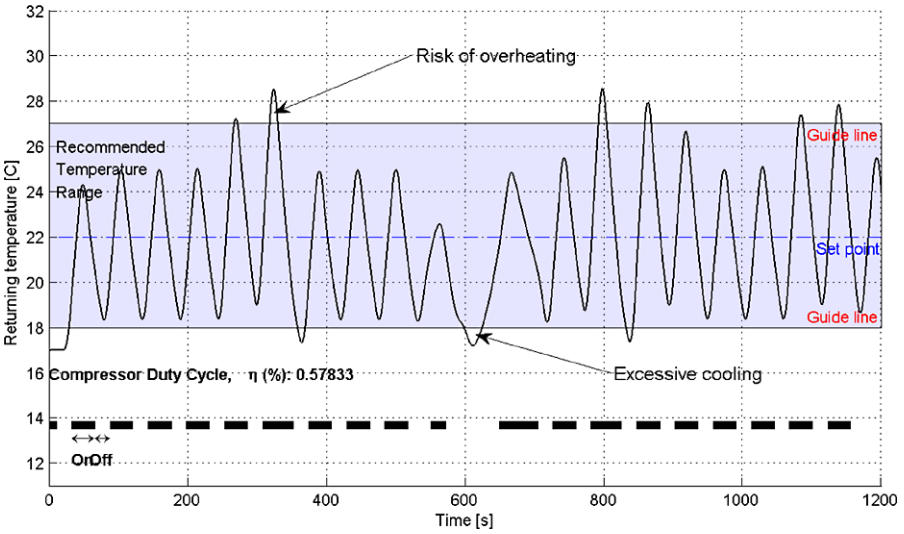

Fig. 9 Reactive Approach: Compressor cycle change vs. returning temperature of the CRAC; Positive heat imbalance creates "risk of overheating" and negative heat imbalance shows "excessive cooling" when fan speed ω is fixed as 5 m^3/s

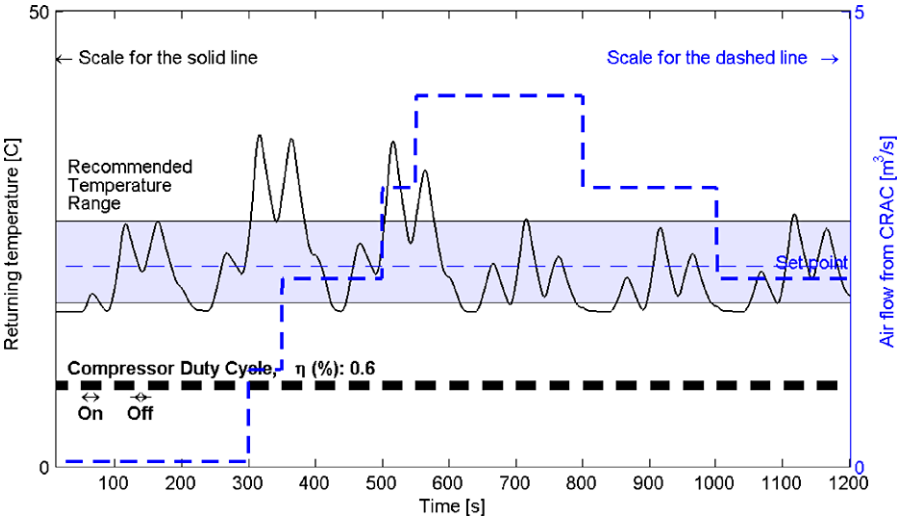

Fig. 10 Reactive Approach: Fan speed change vs. returning temperature of the CRAC; "risk of overheating" still exist even if the CRAC fan speed increases. Duty cycle of the air compressor η is fixed as 0.6

set point to react based on the temperature recommended in [2], which ranges from 64.4°F (18°C) to 80.6°F (27°C). We change the fan speed or duty cycle of the compressor to adjust temperature to the set point as shown in Figs. 9 and 10. Since controller does not have knowledge about how much heat will be generated in the future,

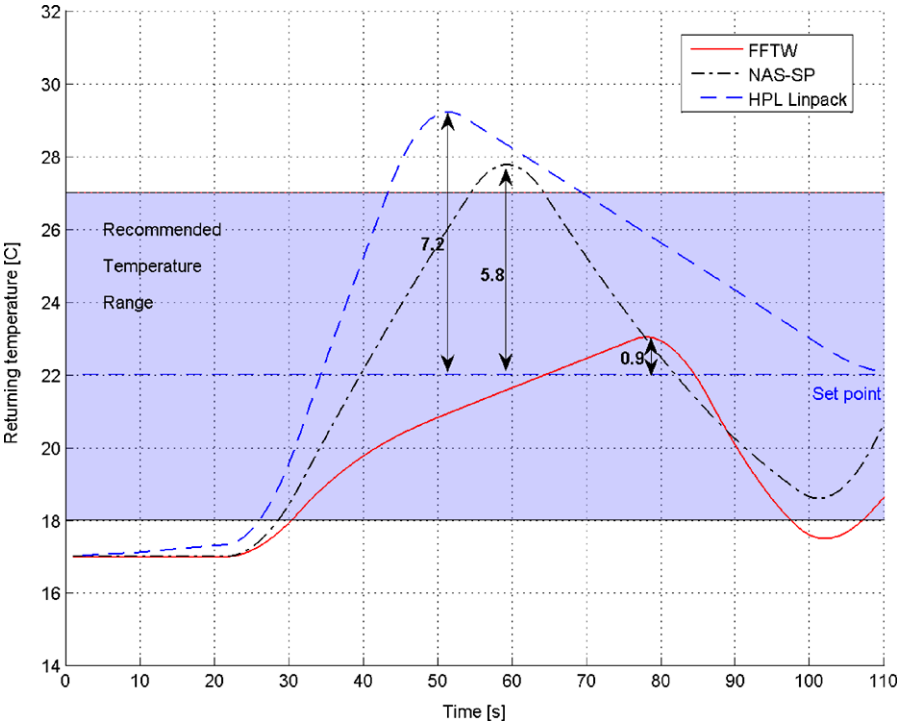

Fig. 11 Temperature variations for FFTW, NAS-SP and HPL workloads

it can only adjust fan speed or duty cycle based on the temperature variations in this approach.

The reactive approach takes a corrective action after the temperature has crossed a set point temperature. Note that this set point temperature can be different based on the response time of the control system used. As a reaction, duty cycle of the compressor can vary based on the changes in returning air temperature as shown in Fig. 9. If returning temperature increases above the set point, then the controller increases duty cycle of the compressor accordingly to extract more heat and hence lower the temperature. Fan speed is fixed at its maximum value which is 5 m^3/s. However, the "risk of overheating," which refers to the amount of heat that can damage internal components, exists because of the noninstantaneous cooling effect on the blades. Also, because of the delayed reaction of the cooling system, excessive amount of heat may be extracted (excessive cooling) bringing the temperature below the guide line.

Changing the fan speed to adjust the temperature is the alternative way to control the temperature in reactive approach. Figure 10 shows CRAC fan speed control and corresponding returning temperature when compressor cycle is fixed at 0.6 to show the effect of fan in this case. Fan speed is controlled by the controller as the temperature increases above the set point or decreases below the set point, but the "risk of overheating" and "excessive cooling" still exists because of the delay.

We observed the temperature variations for different workloads such as FFTW, NAS-SP, and HPL. From Fig. 11, we see that the temperature change for different

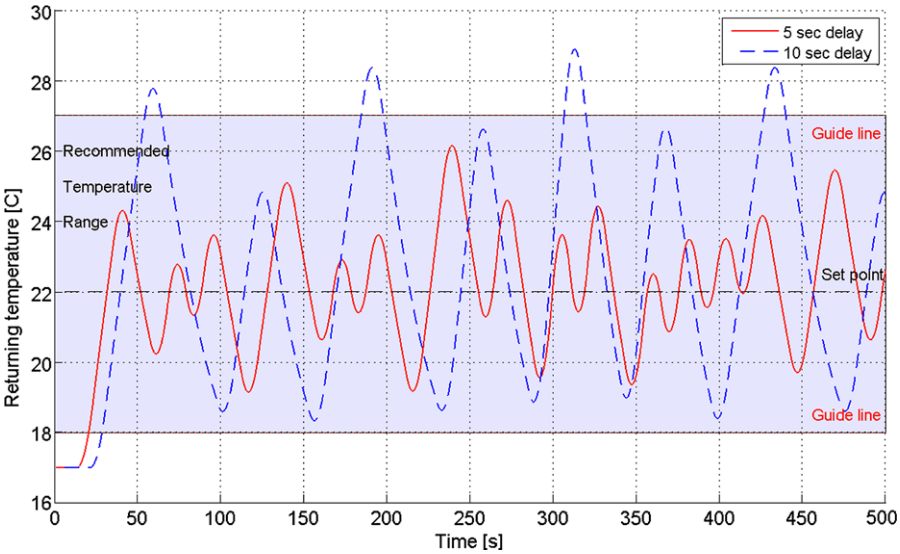

Fig. 12 Temperature variations based on the time delay which is dependent on the distance of the blade from the CRAC unit

workload is different. From this, we infer that the temperature variation is dependent on the workload and each one of them has a different heat pattern. In Fig. 11, it is easy to control the temperature for FFTW workload as it generates less heat compare to the other two. In the case of HPL, it is difficult to control the temperature because it generates more heat and the temperature rises quickly above the recommended temperature range.

In Fig. 12, we show the temperature variations based on the time delay which is dependent on the distance of the blade from the CRAC unit. As the heat propagation is dependent on the distance of the heat sensors from the source of the heat generation, there is a delay in detection of the rise in temperature at the CRAC unit. Hence, there is a delay in the response of the CRAC unit to control the temperature at the source of heat generation. In Fig. 12, we see that higher the delay, more is the chance of temperature rising above recommended temperature range.

5.2 Proactive approach

In proactive approach, "duty cycle of a compressor (η)" or "fan speed of the CRAC (ω)" can be jointly optimized upon the heat estimation model provided in Sect. 4. Optimization problem is solved every 50 s, which is time window size of air compressor duty cycle (T_{ON} + T_{OFF}) to adjust η and ω. This model is evaluated for the three chosen benchmarks. Proactive approach is intrinsically predictive in nature as it estimates the heat that will be generated in the future and adjusts CRAC unit accordingly. This way, we can prevent "risk of overheating" and "excessive cooling" by eliminating the delay in cooling action. In Fig. 13, we plot the percentage of server blade utilization with respect to time. It provides us the information about the work-

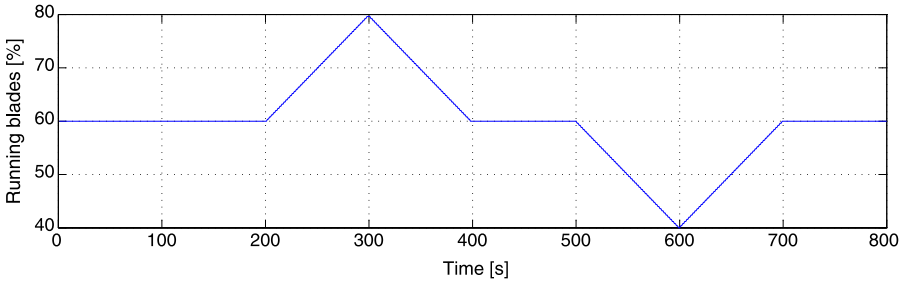

Fig. 13 Utilization rate of server blades [%]

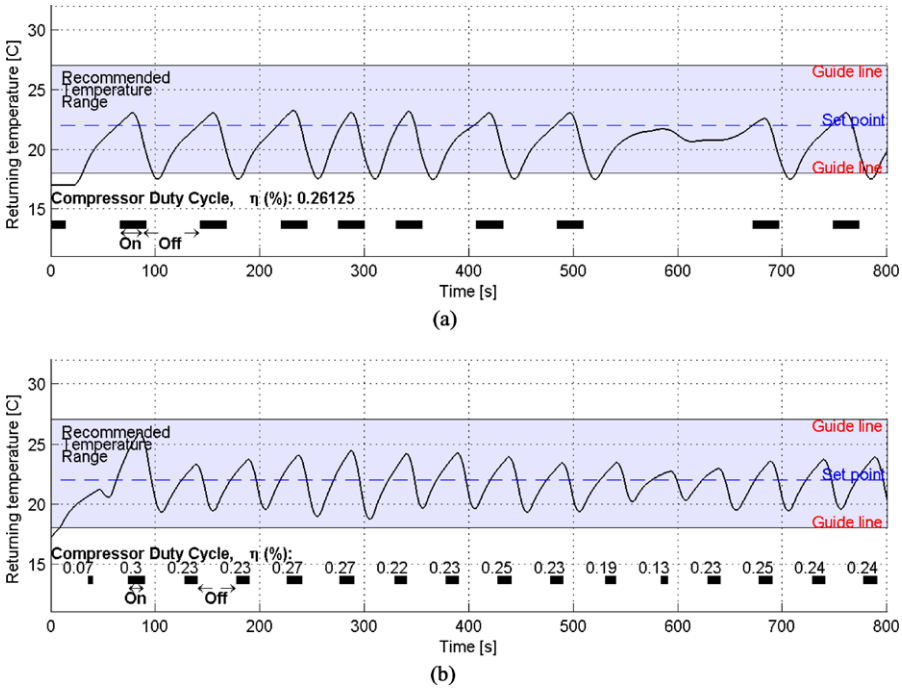

Fig. 14 Temperature variation of datacenter, FFTW workload, Random load balancing for the job distribution (a) Reactive approach; (b) Proactive approach

load intensity at a particular time instance. This server utilization rate is same for all the other workload simulations.

Figures 14(a) and 14(b) show the temperature changes based on FFTW workload subsystem usage profile. Workloads are assigned to the blades by random load balancing algorithm. Compressor cycle in reactive approach in Fig. 14(a) decreases when the temperature crosses the set point. This workload shows moderate temperature change because this workload does not generate much heat compared to the other two workloads. In this case, "excessive cooling" appears due to noninstantaneous action of air conditioning system. Figure 14(b) shows that temperature remains around

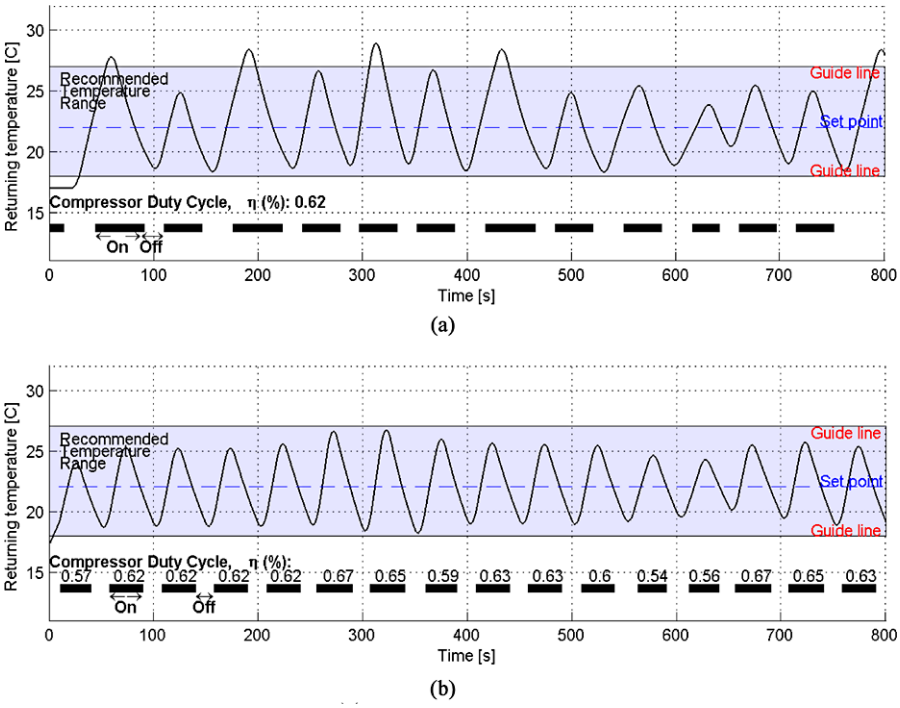

Fig. 15 Temperature variation of datacenter, NAS-SP workload, Random load balancing for the job distribution (**a**) Reactive approach; (**b**) Proactive approach

the "set point," and in turn saves energy by optimizing fan speed and compressor cycle. Energy consumption is compared in Sect. 5.3.

Figures 15(a) and 15(b) show the temperature change based on NAS-SP workload subsystem usage profile. In Fig. 15(a), we observe a periodic "risk of overheating" during the time of simulation due to delay in cooling. This approach causes almost a 10-second delay because of the distance of the blade from the CRAC unit. On the contrary, proactive approach estimates the heat to be generated, and optimizes fan speed and compressor cycle based on the estimation by the model proposed in Sect. 4. Figure 15(b) shows that temperature varies mostly within the recommended temperature range.

Figures 16(a) and 16(b) show the temperature change based on HPL-Linpack workload subsystem usage profile that has a highest heat generation among the three benchmarks. Since reactive approach (Fig. 16(a)) does not have any knowledge about the workload, the controller increases the compressor cycle and fan speed according to increase in temperature. A high "risk of overheating" appears between 250 to 350 seconds because heat imbalance (Q) is positive and very high. The compressor is too slow to react to the temperature rise because HPL generates more heat as compared to other workloads. Also, from Fig. 13, we see that the highest blades utilization occurs in this same period of time. Due to this, there is extreme load on the CRAC, and hence the slow reaction. On the contrary, proactive approach optimizes fan speed and

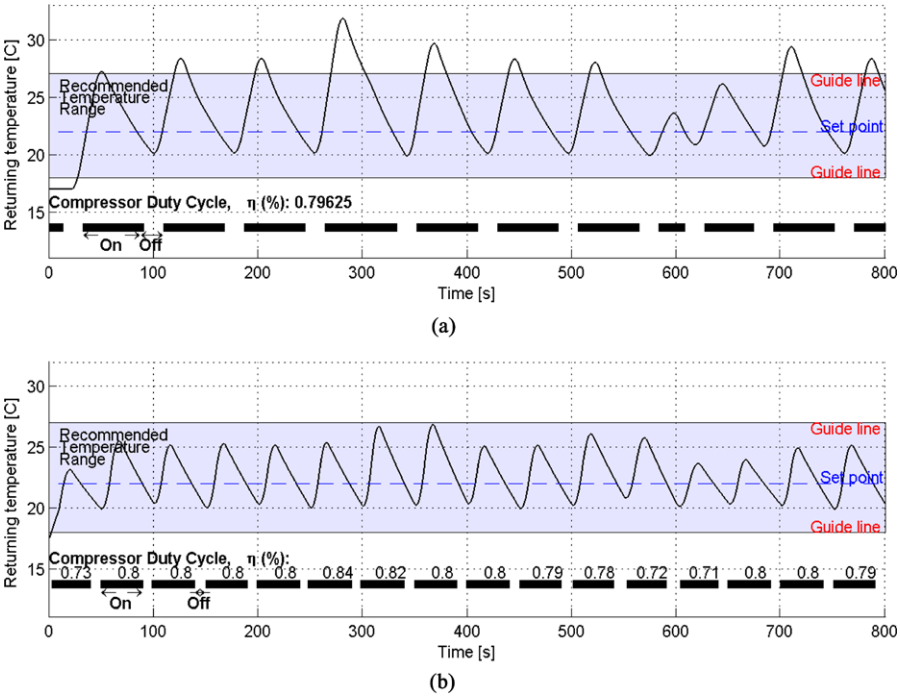

Fig. 16 Temperature variation of datacenter, HPL Linpack workload, Random load balancing for the job distribution (**a**) Reactive approach; (**b**) Proactive approach

compressor cycle based on the estimation of heat to be generated. Figure 16(b) shows that temperature varies mostly within the recommended temperature range.

The heat generation in a datacenter is dependent on workload distribution, with uneven distribution of the workload there is uneven heat generation. With a global control system and a single fan, we cannot partially extract the heat. Ideally one fan can extract the same amount of heat from each blade, but cannot extract heat from a selected blade, enclosure or a rack. Moreover, selective cooling using multiple fans is needed for an intensive workload since, it can causes a high heat imbalance and isolated hot spots. A selective air conditioning system control using multiple fans (one per "corridor" j) is an energy efficient alternative, since we increase the fan speed only when needed depending on uneven heat generation.

5.3 Energy consumption and overheating risk

We estimate the "energy consumption" for cooling systems and the "risk" of overheating for the hardware components, which is the input to both reactive approach and proactive approach models. Energy consumption is the sum of energy consumed by the fan and the air compressor during the time for which the simulation runs. The energy is represented in the units of "kWh," which can be directly converted into kJ, multiplying by a factor of 3600. "Risk" refers to percentage of overheating risk of the hardware components, calculated by averaging the percentage of blades over the

threshold which is set as 70° Celsius in this simulation. In this way, we can show what percentage of blades are under the state of overheating risk in a datacenter.

Tables 4 and 5 show the results for random load balancing and sequential load balancing, respectively. Simulations are performed by using parameters in Table 3. Table 5 shows lower energy consumption than Table 4 because in Table 4 workloads are evenly distributed in an orderly manner so that the heat generation is uniform.

We compare the proactive and reactive approach based on their energy consumption and "risk" factor. In reactive approach, using return temperature (global temperature measurement) for activating CRAC unit shows lower energy consumption than using internal temperature (local temperature measurement) because it only uses mixed returning air temperature that averages heat imbalance of all the blades in a datacenter. However, using the internal temperature of the blade activates CRAC unit whenever the temperature at any blade crosses threshold, and therefore prevents the "risk" of overheating, but while avoiding this risk consumes more energy compared to the approach that uses return temperature (global temperature measurement).

The proactive approach does not have big "risk" compared to the reactive approach since the controller can quantify and extract heat before it creates heat imbalance. However, the proactive approach is inherently based on localized estimation of heat generated at each blade and hence, it consumes more energy compare to the reactive approach using single fan in Tables 4 and 5.

In proactive approach, using multiple fans consumes less energy than using single fan or reactive approach. Even though multiple fans require additional power to operate which is few kilowatts, they can adjust the airflows to different aisles and selectively extract heat that causes the 'risk' of overheating efficiently. Since proactive approach is by default based on localized estimation, multiple fans help remove this heat in a localized manner, which results in increase in the energy efficiency. Difference in energy consumption between using multiple fans and single fan is more apparent in sequential load balancing in Table 5. Using multiple fans (Fig. 17(b)) achieves lower energy consumption than using single fan (Fig. 17(a)), showing that temperature changes in lower range using lower compressor duty cycle.

6 Conclusions and future work

In this paper, we proposed a proactive approach based optimization model for cooling systems of the datacenters. The proactive approach is based on having advanced knowledge of the workload behavior and taking an appropriate action before the heat imbalance affects the temperature. When we compared the proactive approach with the reactive approach, reactive approach was found to have many disadvantages such as delayed response, high risk of over heating, excessive cooling, and recursive cycling. Proactive approach proposed in this paper overcomes these disadvantages of the reactive approach. Proactive approach cools the system before temperature rises and prevents any occurrence of "risk of overheating" and also prevents "excessive cooling" as the heat imbalance is estimated based on the knowledge of subsystem usage and the workloads. For it to be effective there is a need for multiples fans under each plenum for each aisle or each corridor, which is possible with minimal or no

Table 4 Risk of overheating and Energy consumption for random load balancing

Workload Type	Reactive				Proactive			
	Return temperature (Global Temperature Msmt.)		Internal temperature (Local Temperature Msmt.)		Single fan (Global control of the fan)		Multiple fans (Local control of the fan)	
	Risk (%)	Energy (kWh)	Risk	Energy	Risk	Energy	Risk	Energy
FFTW	0.00	7.61	0.00	7.92	0.00	8.23	0.00	7.27
NAS	3.27	14.84	0.58	15.45	0.03	15.03	0.05	14.22
HPL	15.55	18.12	16.35	20.09	3.10	18.515	3.71	16.19

Table 5 Risk of overheating and Energy consumption for sequential load balancing

Workload Type	Reactive				Proactive			
	Return temperature (Global Temperature Msmt.)		Internal temperature (Local Temperature Msmt.)		Single fan (Global control of the fan)		Multiple fans (Local control of the fan)	
	Risk (%)	Energy (kWh)	Risk	Energy	Risk	Energy	Risk	Energy
FFTW	0.00	7.62	0.00	7.93	0.00	7.99	0.00	7.12
NAS	2.74	13.52	0.36	14.57	0.28	14.24	0.00	13.83
HPL	14.65	18.74	11.02	19.15	2.55	20.95	2.44	17.57

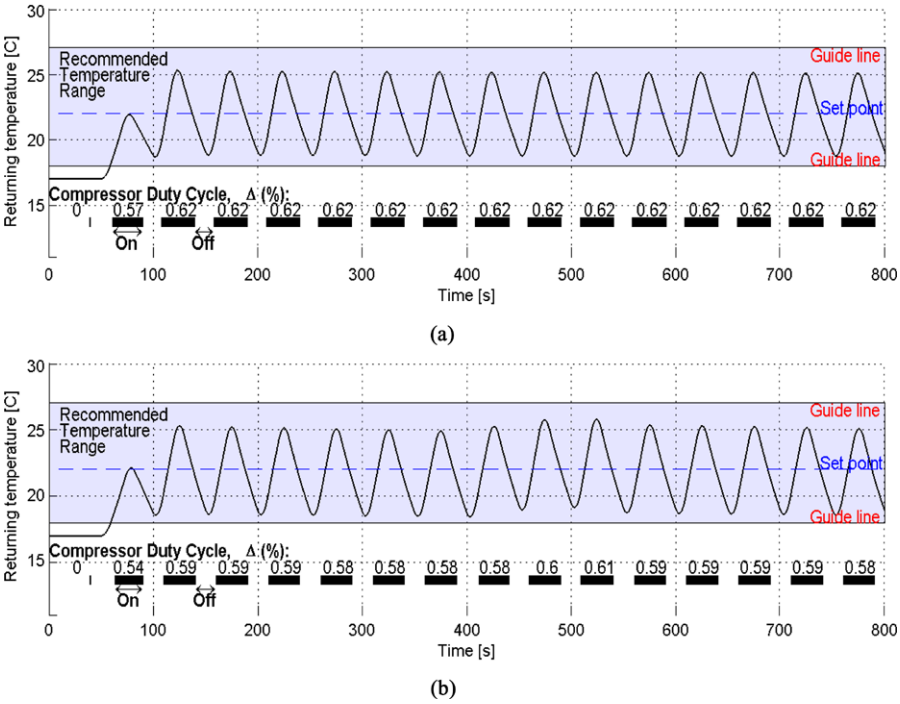

Fig. 17 Temperature variation of the datacenter, NAS-SP workload, Random load balancing for the job distribution (**a**) Proactive approach with single fan; (**b**) Proactive approach with multiple fan

changes to the designs of the existing datacenters. Multiple fans help in effectively controlling hot spots occurring in different locations, which cannot be eliminated by the current single fan cooling systems even by using proactive approach. The use of multiple fans in the proactive approach to control the cooling system saves approximately 4% to 10% of the energy required for cooling, depending on the workload and scheduling algorithms implemented.

Proactive approach suggested in this paper can be optimized if it is implemented with supporting infrastructures such as external temperature and humidity sensors and airflow meters that are crucial in obtaining inlet and outlet temperatures, humidity and airflow for accurate air circulation modeling in (8). Thermal cameras could be another option to micro-managing heat imbalance and hot spots. Thermal cameras can be used to detect, characterize, localize, and track hot spots causing heat imbalances. Future work involves implementing proactive approach using this infrastructure and analyzing the improvement in the performance obtained.

References

1. Advanced configuration & power interface (acpi). Technical report, 2009
2. ASHRAE Technical Committees (2004) Thermal guidelines for data processing environments. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE)

3. Beitelmal AH, Patel CD (2007) Computational fluid dynamics modeling of high compute density data centers to assure system inlet air specifications. Distrib Parallel Databases 21(2–3):227–238
4. Cameron KW, Ge R, Feng X (2005) High-performance, power-aware distributed computing for scientific applications. Computer 38(11):40–47
5. Chandra G, Kapur P, Saraswat KC (2002) Scaling trends for the on chip power dissipation. In: Proc of IEEE interconnect technology conference (IITC), Burlingame, CA, June 2002, pp 170–172
6. EPA (2007) EPA report to congress on server and data center energy efficiency. Technical report, US Environmental Protection Agency
7. Greenberg S, Mills E, Tschudi B (2006) Best practices for data centers: lessons learned from benchmarking 22 data centers. In: Proc of American council for an energy-efficient economy (ACEEE), Pacific Grove, CA, August 2006
8. Hayama H, Nakao M (1989) Air flow systems for telecommunications equipment rooms. In: International telecommunications energy conference (INTELEC), Florence, Italy, October 1989
9. Hayama H, Nakao M (1990) Airflow distribution in telecommunications equipment rooms. In: International telecommunications energy conference (INTELEC), Orlando, FL, October 1990
10. Heath T, Centeno AP, George P, Ramos L, Jaluria Y, Bianchini R (2006) Mercury and freon: temperature emulation and management for server systems. In: Proc of the architectural support for programming languages and operating systems (ASPLOS-XII), San Jose, California, October 2006, pp 106–116
11. Hsu C, Feng Wu, Archuleta JS (2005) Towards efficient supercomputing: a quest for the right metric. In: Proc of IEEE international parallel and distributed processing symposium (IPDPS), Denver, Colorado, April 2005
12. Kang S, Schmidt RR, Kelkar K, Patankar S (2001) A methodology for the design of perforated tiles in raised floor data centers using computational flow analysis. IEEE Trans Compon Packag Technol 24(2):177–183
13. Khan SU, Ahmad I (2009) A cooperative game theoretical technique for joint optimization of energy consumption and response time in computational grids. IEEE Trans Parallel Distrib Syst 20(3):346–360
14. Liu Y, Zhu H (2009) A survey of the research on power management techniques for high-performance systems. Softw Pract Exp
15. Moore J, Chase J, Ranganathan P, Sharma R (2005) Making scheduling "cool": temperature-aware workload placement in data centers. In: Proc of annual conference on USENIX annual technical conference (ATEC), Anaheim, CA, April 2005, p 5
16. Mukherjee T, Banerjee A, Varsamopoulos G, Gupta SKS, Rungta S (2009) Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. Comput Netw 53(17):2888–2904
17. Nakao M, Hayama H, Nishioka M (1991) Which cooling air supply system is better for a high heat density room: underfloor or overhead? In: Proc of international telecommunications energy conference (INTELEC), Kyoto, Japan, November 1991
18. Patel C, Bash C, Belady L, Stahl L, Sullivan D (2001) Computational fluid dynamics modeling of high compute density data centers to assure system inlet air specifications. In: Proc of pacific Rim/ASME international electronic packaging technical conference of (IPACK), Kauai, HI, August 2001
19. PGnE (2006) High performance data centers: a design guidelines sourcebook. http://hightech.lbl.gov/documents/data_centers/06_datacenters-pge.pdf
20. Rambo J, Joshi Y (2007) Modeling of data center airflow and heat transfer: state of the art and future trends. Distrib Parallel Databases 21(2–3):193–225
21. Schmidt RR (2004) Thermal profile of a high-density data center-methodology to thermally characterize a data center. Trans Am Soc Heat Refrig Air-Cond Eng (ASHRAE) 110(2):635–642
22. Schmidt RR, Cruz E (2002) Raised floor computer data center: effect on rack inlet temperatures of exiting both the hot and cold aisle. In: Proc of intersociety conference on thermal phenomena in electronic systems (ITHERM), San Diego, CA, August 2002
23. Schmidt RR, Karki K, Kelkar K, Radmehr A, Patankar S (2001) Measurements and predictions of the flow distribution through perforated tiles in raised floor data centers. In: Proc of pacific Rim/ASME international electronic packaging technical conference of (IPACK), Kauai, HI, August 2001
24. Schmidt RR, Karki KC, Patankar SV (2004) Raised floor computer data center: perforated tile flow rates for various tile layouts. In: Proc of intersociety conference on thermal phenomena in electronic systems (ITHERM), Las Vegas, NV, June 2004
25. Schmidt RR, Cruz EE, Iyengar MK (2005) Challenges of data center thermal management. IBM J Res Dev 49(4/5):709–723

26. Shan AJ, Krishnan N (1989) Flow resistance: a design guide for engineers. Hemisphere, Washington
27. Sharma RK, Bash CE, Patel RD (2002) Dimensionless parameters for evaluation of thermal design and performance of large-scale data centers. In: Proc of ASME/AIAA joint thermophysics and heat transfer conference, St. Louis, MO, June 2002
28. Srinivasan J, Adve SV, Bose P, Rivers JA (2004) The impact of technology scaling on lifetime reliability. In: Proc of the international conference on dependable systems and networks (DSN), Washington, DC, 2004. IEEE Computer Society, Washington, p 177
29. Subrata R, Zomaya AY, Landfeldt B (2010) Cooperative power-aware scheduling in grid computing environments. J Parallel Distrib Comput 70(2):84–91