Preproceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 21–34

Machine Learning in Energy and Thermal-aware Resource Management of Cloud Data Centers: A Taxonomy and Future Directions

Shashikant Ilager[1,2] and Rajkumar Buyya[1]

[1]Cloud Computing and Distributed Systems (CLOUDS) Lab
School of Computing and Information Systems
University of Melbourne, Australia

[2]Institute of Information Systems Engineering
TU Wien, Austria

***Abstract*—Cloud data centres (CDCs) are the backbone infrastructures of modern digital society, but they also consume huge amounts of energy and generate heat. To manage CDC resources efficiently, we must consider the complex interactions between diverse workloads and data centre components. However, most existing resource management systems rely on simple and static rules that fail to capture these complex interactions. Therefore, we require new data-driven Machine learning-based resource management approaches that can efficiently capture the interdependencies between parameters and guide resource management systems. This review describes the in-depth analysis of the existing resource management approaches in CDCs for energy and thermal efficiency. It mainly focuses on learning-based resource management systems in data centres and also identifies the need for integrated computing and cooling systems management. A taxonomy on energy and thermal efficient resource management in data centres is proposed. Furthermore, based on this taxonomy, existing resource management approaches from server level, data centre level, and cooling system level are discussed. Finally, key future research directions for sustainable Cloud computing services are proposed.**

***Index Terms*—Cloud Computing, Energy Efficiency, Thermal-aware Workload Management, Sustainable Computing, Machine Learning**

I. Introduction

CLOUD computing has changed the way computing services are delivered to end-users by providing flexible and on-demand access to resources with a pay-as-you-go model [1], [2]. Cloud computing follows the principle of providing computing resources as utility services (e.g., water and electricity). This unique and flexible service delivery model ensures that individuals and businesses can easily access the required computing services. By default, Cloud workloads require continuous, always-on, and 24×7 access to its deployed services. For instance, the Google search engine is expected to achieve an almost 100% availability rate [3]. Similarly, Amazon AWS witnesses thousands of Elastic Compute (EC2) instances created [4] in a day through their automated APIs, thus requiring CDCs to support such critical demand. According to Gartner, by 2022, 60% of organisations will use an external Cloud service provider [5], and by 2024, Cloud computing alone will account for 14.2% of total global IT spending [6].

Cloud computing services are broadly categorised into three types. First, the Infrastructure as a Service (IaaS) model offers computing, storage, and networking resources either in virtual or physical form. Second, the Platform as a Service (PaaS) model offers tools for rapid application development and deployment, such as middleware platforms, Application Programming Interfaces (APIs), and Software Development Kits (SDKs). Third, the Software as a Service (SaaS) model offers direct access to application software to the users, and the software is developed and managed by service providers completely.

All of these service paradigms rely on the data centres to deliver the resources required for the applications and users seamlessly. Cloud Data Centres (CDCs) are massive network-based infrastructures managed in runtime by Resource Management Systems (RMS). Fig. 1 shows an abstract view of data centre infrastructure and its resource management system. The DCs host thousands of servers, networking equipment, and cooling systems. Servers and networking equipment provide the required computational resources for cloud users, and the cooling system helps to remove the heat generated by the computing resources. An RMS in the data centre is a software platform that manages different subsystems in the data centre through various tasks, such as resource monitoring, provisioning, workload scheduling, and placement. It also controls power and cooling management knobs. Some public cloud service providers build their own in-house RMS, while many private and public clouds use open-source systems such as OpenStack[1].

Note: This work is done when first author was working at University of Melbourne, Australia

[1]https://www.openstack.org

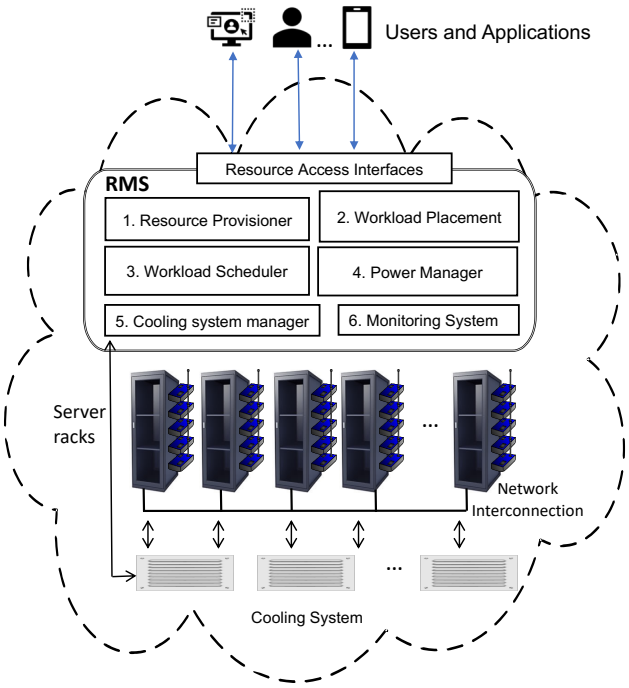

Fig. 1: An abstract view of a CDC. It shows high level view of an RMS tasks required to manage the CDCs resources and user workloads (adapted from [7]).

To meet the demand for Cloud services, major Cloud services providers such as Amazon AWS[2], Microsoft Azure[3], and Google Cloud[4] are deploying many hyper-scale data centres in multiple regions worldwide. There are over 8 million data centres globally, ranging from private small-scale to hypers-scale DCs, and they are growing at 12% annually [8]. As they grow in number and size, they consume more energy and face massive energy challenges. CDCs consume an estimated 2% of global electricity generated [9] and rely on fossil-fuel-based or brown energy sources that emit 43 million tons of CO_2 per year and increase at 11% annually [10], leaving high carbon footprints. Therefore, improving the energy efficiency of Cloud data centres is vital for sustainable and cost-effective Cloud computing. DCs' tremendous growth has introduced massive energy challenges. If not addressed, data centres may consume up to 8000 terawatts of power by 2030 in the worst case. However, if best practices are adopted across the Cloud computing stack, this energy consumption can be reduced to around 1200 terawatts [11] (see Fig. 2). To achieve this best-case scenario, energy-efficient practices are needed in various levels of data centre resource management platforms (such as optimised use of computing and cooling resources). Hence, addressing this energy problem and achieving sustainability, both environmentally and economically, is crucial.

[2]https://aws.amazon.com/
[3]https://azure.microsoft.com/
[4]https://cloud.google.com/

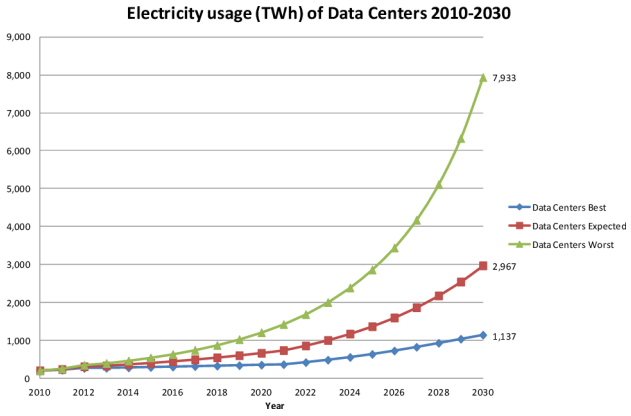

Fig. 2: Estimation of Data Centre Energy Consumption by 2030 [11]

A data centre is a complex cyber-physical system (CPS) that consists of thousands of rack-mounted physical servers, networking equipment, sensors, cooling systems, and other facility-related subsystems. It consumes up to 30-40 kW per rack and generates a lot of heat, posing a serious challenge for efficient and reliable resource and energy management. In particular, the main power consuming subsystems in CDCs are the computing and cooling systems, which together account for 85% of total energy consumption in a data centre, with each of them significantly contributing [12] to the total power consumption. Therefore, there is a essential requirement for integrated Energy and Thermal-aware Resource Management.

Traditionally, cooling system management and computing system management are done separately by the facility management team and the IT administrator, respectively. However, optimising one system may have a negative impact on the other system. For example, increasing resource utilisation in computing may create hotspots, and thus increase cooling energy costs. Therefore, managing these subsystems independently may result in energy inefficiencies in the data centres even if they are individually optimised for energy efficiency. The advancement in IoT and smart systems [13] has enabled many mechanical systems associated with cooling to be controlled or configured through software systems [14]–[16]. Hence, it is crucial to apply resource management techniques holistically to optimise both computing and cooling systems and avoid conflicting trade-offs between these two subsystems.

Resource management in data centres is highly challenging due to the complex interactions between subsystems and the heterogeneous characteristics of workloads. Manual fine-tuning of the controllable parameters by resource management systems is infeasible. For example, *"Just* 10 *pieces of equipment, each with* 10 *settings, would have* 10 *to the* 10^{th} *power, or* 10 *billion possible configurations, a set of possibilities far beyond the ability of anyone to test for real"* [17], [18]. Moreover, these large-scale systems often have nonlinear relationships between their parameters. However, optimising data centre operation requires adjusting

the hundreds of parameters in different subsystems where heuristics or static solutions are ineffective.

Therefore, to cope with the complexity of data centre infrastructures and the dynamic nature of cloud workloads, Machine Learning (ML)-based resource management methods are vital. In parallel, integrated resource management of the computing and cooling systems is necessary to balance the trade-offs between these two subsystems and achieve significant energy efficiency in CDCs [7]. There have been many efforts in this direction using ML for systems focusing on optimising different computing systems [19]. For instance, ML-centric Cloud [20], developed Resource Control (RC), a general ML and prediction serving system that provides insight into the Azure compute fabric resource manager's workload and infrastructure. Similarly, Google has used ML models to optimise the efficiency of its data centres by adjusting the different knobs of the cooling system, thus saving a significant amount of energy [21]. These use cases demonstrate the feasibility and benefits of learning-based solutions in different aspects of resource management in clouds. Moreover, even a 1% improvement in data centre efficiency can save millions of dollars per year and reduce the carbon footprint [22].

The rest of the paper is organised as follows: Section II provides overview of ML-based RMS in CDCs. Section III and Section IV review the existing methods for energy and thermal management in data centres based on the taxonomy, respectively. Section V explains the integrated resource management solutions for energy and thermal efficiency. Section IV-C describes different cooling systems in a data centre, including air and liquid cooling systems. Section VI outlines the future research directions. Finally, Section VII concludes the paper.

II. Background: ML-based Resource Management Systems in CDCs

Machine learning (ML) is naturally used in Computer Vision (CV) and Natural Language Processing (NLP) problems due to its ability to identify patterns from the complex input data. ML algorithms are classified into supervised and unsupervised learning, depending on the input data preparation and training methods. ML methods itself can be broadly used for numerical prediction- *regression models*, and for categorization based on class labels,- *classification modes*, as well for developing advanced control systems- *Reinforcement Learning (RL) controllers*.

As data center complexities increase, ML algorithms are required to perform a variety of RMS tasks. For instance, as illustrated in Figure 3, the left side of the Figure 3 indicates the high level RMS Tasks in a CDC (also see Figure 1). Please note that, these tasks indicate primary functionality of an CDCs middleware system; there could be other tasks based on data center and workload requirements. Similarly, the right side of the Figure indicates list of all possible ML tasks an RMS could require in its decision making process. For example, the *Resource Provisoner* can invoke resource estimation models to predict the required amount of computing

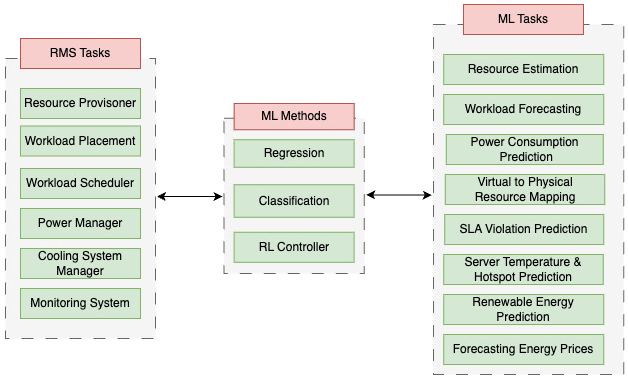

Fig. 3: List of RMS Tasks and ML Tasks and ML Methods.

resources for a workload. The RMS can also be guided by power consumption and SLA violation prediction models, depending on the optimization objectives.

Much advanced ML applications such as RL can be used to develop controller systems. An RMS can be modeled as a decision engine with a list of actions designed to satisfy specific objectives. Such approaches are increasingly being used in CDCs for cooling system knob configuration, scheduling, and power management systems. Although Figure 3 provides an overview of how ML can be leveraged to develop highly optimised RMS systems for today's complex CDCs, it is not exhaustive.

III. Taxonomy of Energy Management in CDCs

Many researchers have focused on increasing the energy efficiency of data centres with various resource management techniques. These techniques cover an individual server to geo-distributed data centres. Taxonomy on the data centre's energy management solutions is presented in Fig. 4. We categorise these solutions into two broad categories, i.e., single server level and data centre level solutions. Accordingly, we identify the essential techniques used in these two categories and briefly review their methods.

A. Server Level

In a computing server, the CPU predominantly consumes a significant amount of energy. Modern rack-mounted data centre servers consume more than 1000 watts of power. Hence, managing this high power consumption is a challenging task. This server-level power management has been mostly left to the operating system and its device drivers that communicate with underlying hardware signals and manage the server power. Server-level power management can be broadly categorized into two levels, static and dynamic power management. Static power management deals with minimising leakage power, while dynamic power management deals with regulating active runtime power based on utilization level.

1) Static Power Management: The silicon chip has static power consumption, which is independent of the usage level. Static power mainly accounts for the leakage of current inside active circuits. To some extent, static power consumption

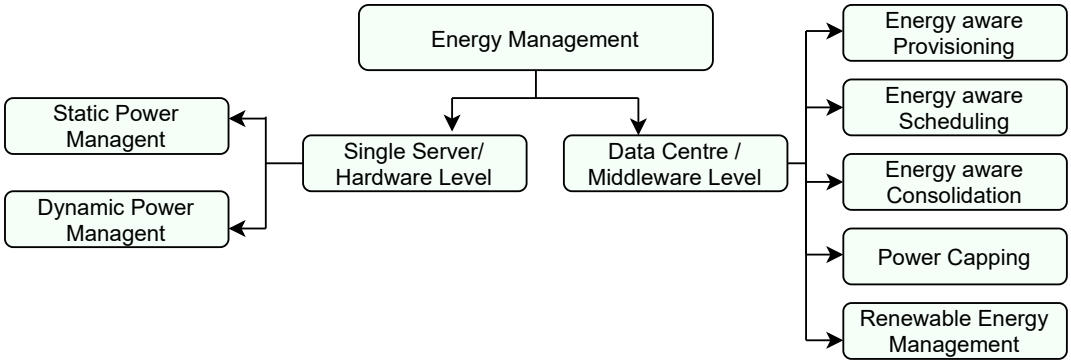

Fig. 4: Taxonomy of Energy Management in Cloud Data Centres

is unavoidable; however, it can be minimized with better design and processes. There are many solutions from a lower level from circuit level and architectural techniques [23]. The general approach in managing leakage is with different sleep states of CPUs when the system is idle. For instance, Intel X86 architecture has (C0-C4) sleep states indicating C0 is an active state, while C4 is a deep sleep state where most of the CPUs' components are turned off to avoid static power consumption. This processor's sleep state management is usually done in reactive manner at Operating System's (OS) kernel level. If a processor core is idle predefined time interval, a kernel governor changes the sleep state.

Application of ML: However, existing reactive static power management approaches could be vastly improved using the ML-based solutions with proactive strategies. For instance, Chung et al. [24] proposed a power management technique for an arbitrary number of sleep states, which turns of idle processors based on idle period clustering and adaptive learning trees. Instead of predefined interval, they estimate the adaptive intervals based on recent history, saving the energy consumption. Similarly, Lu et al. [25] introduce RAMZzz, a memory system design which is based on rank-aware energy-saving optimizations, in memory systems. It groups pages with similar access patterns into the same rank, allowing for dynamic page migrations to optimize access locality and employs adaptive state demotions with a prediction model to increase the energy efficiency. These studies indicates that ML has been widely getting used at very low level management of hardware devices.

2) Dynamic Power Management (DPM): A large part of silicon chip-based computing elements, either in CPU or GPUs spend on dynamic power. Dynamic power represents runtime energy based on workload utilisation level. CPUs operate at different frequencies to regulate the dynamic power. If the operating frequency of a CPU is the highest, then its dynamic power consumption will also be higher. The frequency is regulated based on utilisation level and workload requirements to increase their speedup. Dynamic Voltage Frequency Scaling (DVFS) is a popular technique to regulate the dynamic power in modern systems [26]. The dynamic power can be defined as below:

$$P_{dynamic} \ \alpha \ V^2 F \tag{1}$$

In Equation 1, F is the frequency, and V is the supply voltage to the processor. Based on the frequency, the voltage is regulated, and some frequency ranges usually have a similar. If a CPU should be at its highest speed or frequency should be set to a higher level, thus consuming more power. The operating system scales frequency based on its workload and application demands in runtime.

The DVFS-based optimizations are employed using application metrics, VM-level metrics, or even data-center level utilization metrics [27], [28]. A few studies proposed DVFS techniques at the data center level. These solutions include DVFS-aware VM scheduling, consolidation [27], [28], placement of application based on DVFS capabilities [29], and data centre level task scheduling by synchronizing the frequency scaling among multiple machines [30]. All of these works use heuristic based solutions.

Application of ML: ML-based techniques widely used in DPM optimisations. The Authors in [31] proposed ML-based CPU and GPU DVFS regulator for compute-heavy mobile gaming applications that coordinates and scale frequencies with performance and energy improvements. Similarly, in one of our recent study, we explored how ML-algorithms can help us to dynamically configure GPUs clock frequency based on workload requirements such as deadline [32]. Here, we used popular GPU benchmarks Rodinia and Polybench and collected profiled data which includes hardware level counters. This collected data is further used develop ML regression models to estimate power consumption and execution time across configurable memory and streaming processors frequencies. These models are further used to guide a scheduling algorithm to execute application within predefined deadline with minimal energy consumption.

B. Data Centre Level

A significant amount of energy efficiency can be achieved when data centre-level platforms incorporate energy-efficient resource management policies. Distributed data centre applications span hundreds of machines in geo-distributed data

centres; hence, providing energy efficiency holistically across data centre resources and applications is more feasible and yields better results. In this section, we discuss important techniques for data centre-level energy-efficient solutions.

1) Energy-aware Provisioning: Cloud data centres offer computing resources in terms of Virtual Machines (VMs) or containers. Allocating the required amount of resources for the application need is vital to satisfy the SLAs. However, over-provisioning of resources may yield higher energy consumption and monetary cost to the users, while under-provisioning will potentially violate the SLAs. Many researchers have proposed energy-aware resource provisioning techniques. Authors in [33] investigated energy-aware resource allocation for scientific applications. The proposed system EnReal leverages the dynamic deployment of VMs for energy efficiency. Similarly, Li et al. [34] proposed an iterative algorithm for energy-efficient VM provisioning for application tasks. Beloglazov et al. [35] propose various heuristic algorithms for resource allocation policies for VMs defining architectural principles.

Application of ML: Mehiar et al. [36] offered clustering and prediction-based techniques; they used K-means for workload clustering and stochastic Wiener filter to estimate the workload level of each category and accordingly allocate resources for energy efficiency. Recently Microsoft has proposed Resource Control (RC) [20], where they trained ML models to output predictions like VM lifetime, CPU utilisation, and maximum deployment of VMs. These predictions use various resource management problems for better decision-making, including resource provisioning with the right container size for applications. With increasing availability of data in cloud platforms in regard to user workload behaviours, and usage patterns, ML will be key technique to estimate right amount of computing powers required for user requests.

2) Energy-aware Scheduling: Scheduling is a fundamental and essential task of a resource management system in Cloud data centres. It addresses the following question, given an application or set of VMs (considering the application runs inside these isolated VMs), when and where to place these VMs/applications among available physical machines? This decision depends on several factors, including application start time, finish time, and required SLAs. In addition, workload models, whether an application is a long-running (24×7) web application, or a scientific workflow model of which it's tasks need to be aware of precedence constraints, or applications based on IoT paradigm that is predominantly event-driven. Although one can optimise numerous scheduling parameters, many recent studies have focused on energy optimisation as a priority in Cloud data centre scheduling.

Chen et al. [37] propose energy-efficient scheduling in uncertain Cloud environments. They propose an interval number theory to define uncertainty, and a scheduling architecture manages this uncertainty in task scheduling. The proposed PRS1 scheduling algorithm based on proactive and reactive scheduling methods optimises energy in independent task scheduling. Similarly, Huang et al. [38] investigate energy-efficient scheduling for parallel workflow applications in Cloud. Their EES algorithm tries to slack non-critical jobs to achieve power saving by exploiting the scheduling process's slack room. Energy-efficient scheduling using various heuristics for different application models has been a widely studied topic in literature [39]–[41].

Application of ML: A vast number of study explored application of ML in data centre scheduling. Some solutions rely on predictive models and then use them in scheduling algorithms, while other techniques model scheduling as a complete learning-based problem using Reinforcement learning (RL). Berral et al. [42], adopt ML-based regression techniques to predict CPU load, power, SLAs and then use these in scheduling for better decisions. These solutions still use some level of heuristics with integrated prediction models. However, RL-based scheduling is designed to learn and take action in a data centre environment without explicit heuristics. Cheng et al. [43] proposed DRL-based provisioning and scheduling for application tasks in the data centre.

3) Energy-aware Consolidation: Cloud data centres are designed to handle the peak load to avoid potential SLA violations or overload conditions. Hence, the resources are oversubscribed to manage such an adverse situation. However, this oversubscription leads to resource underutilisation in general. It is estimated that Cloud data centres' utilisation level is around 50% on average. Underutilisation of resources is the main factor in the data centre's energy inefficiency as idle or lower utilised servers consume significant energy (up to 70% [44]). Thus, it is necessary to manage workloads under such oversubscribed and underutilised environments. To that end, consolidation has been a widely used technique to increase energy efficiency. It aims to bring the workloads (VMs and containers) from underutilised servers and consolidate them on fewer servers, thus allowing the remaining servers to be kept in sleep/shut down mode to save energy. Many challenges exist in consolidation, including maintaining VM affinity, avoiding overutilisation, minimising SLA violation, and reducing application downtime due to workload migrations.

Beloglazov et al. [35] proposed various heuristics to consolidate the workload and answer the question, including which VMs to migrate, where to migrate and when to migrate to reduce potential SLA Violation. Many other solutions have broadly focused on energy efficiency along with optimising different parameters (cost reduction, failure management, etc) while consolidating workloads in the data centre [45], [46].

Application of ML: ML-based solutions are predominantly used in consolidation [47], [48]. Hsieh et al. [48] studied VM consolidation to reduce power cost and increase QoS. They predict the utilisation of resources using the Gray-Markov-based model and use the information for consolidation. Similarly, the authors [47] also use prediction for consolidation. They predict memory and network usage and perform consolidation of VMs in a data centre along with CPU. Few researchers have also used RL in energy-aware consolidation [49], [50]. Basu et al. [50] proposed Megh—a system that learns to migrate VMs in the data centre using RL. It proposes the dimensionality reduction technique using

dimensional polynomial space with a sparse basis to minimise the state space in their problem. Their system has shown that it achieves better energy efficiency and cost reduction compared to existing heuristics.

4) Power Capping: Data centres are designed to handle peak power consumption based on the workload and cooling system requirements. Hence, in general, data centres are under-provisioned with power. This power capping on data centre servers restricts the amount of energy available to individual servers even though they can consume their maximum limit, thus providing the required speed for workloads [51]. Managing resources and workload effectively in these power-constrained environments is necessary. It is essential to avoid power inefficiencies in limited power allocated across servers to achieve power proportional computing [52].

In this regard, different power capping mechanisms at the Cloud data centre level are studied. The authors [53] proposed a fast decentralised power capping (DPC) technique to reduce latency and manage power at the individual server. Dynamo [54] is the power management system used by Facebook data centres, which has hierarchical power distribution. The lowest level leaf controller regulates power in a group of servers. This leaf controller, based on a high-bucket-first heuristic, determines the amount of energy to be reduced in each server to meet the power cap limits to which it is constrained. It also considers workload priorities and avoids potential performance degradation due to its power capping. Controlling peak power consumption is also a widely studied approach [55] by designing a feedback controller, which periodically reads system-level power and configures the highest power state of servers, keeping the server within its power budget. Authors in [56] studied optimal power allocation in servers, which accounts for several factors, including power-to-frequency, the arrival rate of jobs, and maximum and minimum server frequency configuration. They have shown that allocating full power may not always result in the highest speed as expected. Some techniques have also explored enabling data centre service providers to dynamically manage the power caps by participating in an open electricity market and achieve cost and energy efficiency [57].

Application of ML: Kumbhare et al. [58] propose a prediction based power over subscription in cloud data centres. they predictions of workload performance criticality and virtual machine (VM) resource utilisation and use this information to over subscribe the resources and increase overall utilisation. With Random Forest (RF) and Gradient Boosting (GB) models are used to predict the workload criticality and VM utilisation and use per VM power capping controller to limit its resource usage based on these predictions. However, due to the close interconnection between power capping effect on CPU speed, thermal dissipation and also the presence of heterogeneity in servers and workloads, data centre level power capping workload management is a difficult task to achieve [59] as compared to other energy efficiency methods that are discussed in this paper.

5) Renewable Energy Management: Data centres consume colossal energy and contribute significantly to greenhouse gas emissions (CO_2). Data centre service providers continuously increase renewable or green energy (solar, wind) usage with minimal carbon footprints to decarbonise the data centres. However, green energy usage in the data centre is extremely challenging due to its intermittent nature of availability. In contrast, the Cloud data centre needs an uninterrupted power supply since Cloud workloads tend to run 24×7. Therefore, managing workloads under the uncertain availability of renewable energy is a challenging research problem.

Several resource management techniques explored maximising renewable energy in data centres. They include workload shifting and placement across geo-distributed data centres [60]–[62] based on their carbon efficiency. Besides, delaying job execution if an application can tolerate the QoS [63] and job dispatching or load balancing workloads to match the available renewable energy at different data centres [64] are some popular techniques in this regard.

Application of ML: ML-based algorithms are promising in renewable energy management, as predicting the available green energy based on an environmental condition is crucial in resource management tasks [65]–[67]. For instance, researchers from Google developed [68] carbon aware workload scheduling strategies for batch processing jobs by predicting the next day's available renewable energy from their energy sources. Similarly, Authors in [69] explores forecasting the carbon intensity of geographically distributed data centres and provides temporal shifting of workloads to minizime overall carbon footprints of workload execution. Along with prediction models, RL methods are also used to solve optimisation problems in increasing green energy usages in data centres by mapping renewable energy sources and physical machines [66].

IV. Taxonomy of Thermal Management in CDCs

Similar to energy management, thermal management techniques span from an individual server to data centres. A taxonomy on thermal management solutions is presented in Fig. 5. This section categorises these techniques into two broad categories, i.e., micro-level or single server level and macro-level or data centre-level thermal management techniques. We describe and review existing approaches and ML-based approaches used in these two categories.

A. *Server Level*

To achieve optimal performance, especially in modern chips with very high power densities, thermal constraints are the most critical challenges. Hence, it is essential to operate processors within the predefined Thermal Design Power (TDP) limit [70]. The servers consume an enormous amount of energy and dissipate it as heat. It is crucial to keep processor or CPU temperature within the TDP limit to avoid damage to the processor's silicon components, and prevent from catastrophic device failures. Modern rack servers reach peak temperatures up to 90-100°C.

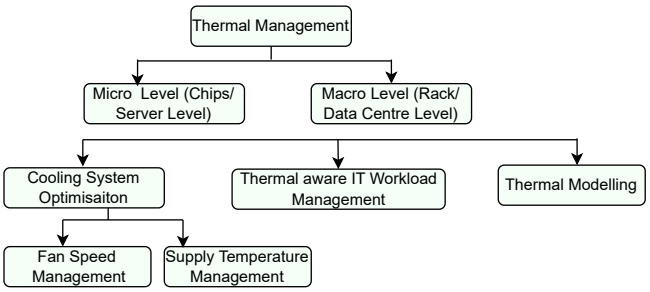

Fig. 5: Taxonomy of Thermal Management in Cloud Data Centres

Like DVFS in energy management, its corresponding thermal dissipation is regulated in servers by controlling the amount of power consumed. Dynamic Thermal Management (DTM) [71] is a popular thermal management technique at the individual server level which regulates Multiprocessors Systems-on-chip (MPSoCs) power consumption, and performance. This is done at the operating system level by closely communicating with underlying hardware interfaces. If a server's temperature is potentially exceeding the predefined TDP, the operating system takes actions by employing thermal throttling mechanisms to reduce energy consumption, thus reducing the CPU speed. Moreover, techniques like dynamic application scheduling [72], [73], onboard fan speed configuration [74] can also be employed for energy and thermal efficiency at the server level.

Application of ML: Recently, ML-based solutions have been applied to optimise temperature management at the individual server level. For example, Iranfar et al. [75] investigated how to proactively estimate the required number of active cores, operating frequency, and fan speed. Accordingly, the system is configured to achieve reduced power consumption and thus regulating corresponding server temperature. Although power consumption and CPU temperature are highly correlated, many other factors affect the thermal behaviour of servers including OS-level scheduling policies and compute-heavy applications. Therefore, analysing such resource management policies and workload behaviour through profiling, bench-marking, and then modelling through ML [32], is crucial for the design of future operating systems. As our focus is entirely on cloud data centres, we delve into more details on data centre level solutions in this regard.

B. Data Centre Level

A typical large-scale CDCs hosts thousands of servers. CDCs servers are arranged in rack-layout, where each rack (e.g., standard 42U rack) can accommodate 10-40 blade servers based on vendor-specific dimensions. This high density of equipment makes the data centre one of the highest-energy-density physical infrastructures. Dissipated heat from these rack servers can result in the data centre's ambient temperature reaching extremely high. Thus, cooling systems in data centres make sure that the data centre temperature is within the threshold. Many approaches exist, optimising different parameters to reduce cooling energy. In this section, we review and describe data centre-level thermal management techniques.

1) Cooling System Configurations: Traditional rack layout data centres have a Computer Room Air Conditioning (CRAC) cooling system that blows cold air to the racks across the data centre (more details of cooling technologies can be found in Section IV-C). The entire cooling system efficiency requires multiple parameters to be configured in the design and operational phase. In the design phase, efficiency can be increased by better physical layout and vent designs to reduce heat re-circulations. While runtime cooling energy efficiency can be increased by fine-tuning the fan speeds of CRAC systems and cold air supply temperature, which determines the cooling system energy consumption [76]–[78]. In this section, we focus on runtime cooling system optimisation.

Fan Speed Management: Within the CRAC system, fans are used to regulate the airflow rate within the data centre. It is important to note that these fan speeds are separate from the onboard server's fan equipped to eject heat from CPU to the outside of the server cabinet. Increasing airflow requires higher fan speeds, thus consuming more energy. Hence, regulating fan speed optimally can save a significant amount of cooling power. However, this depends on the status of the data centre and its temperature level. Many researchers have proposed solutions to optimally configure the CRAC's fan speed based on cooling load [76], [79] by monitoring thermal load in the data centre and accordingly varying fan speeds dynamically to reduce energy consumption.

Supply Temperature Management: CRAC system blows cold air to racks through vented floor tiles in the data centre to take out dissipated heat. Passing colder air requires higher energy consumption as chillers in CRAC consume energy to supply cold air. Hence, the inaccurate configuration of supply air temperature significantly affects cooling energy costs in the data centre. For a safer operation, the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) [80], recommends supply air temperature in the data centre to be in the range of 17-27 °C. Thus, it is beneficial to set the supply temperature closer to 27 °C. However, most data centres are overcooled as the supply temperature in the data centre is set to a much lower temperature conservatively, leaving energy inefficiencies in the cooling system. Setting a higher supply air temperature requires careful handling of peak temperature in data centres.

Many solutions have been proposed to raise the supply air temperature. Zhou et al. [81] have shown that significant power saving can be achieved when the workload is managed efficiently and allows the supply air temperature to be increased. In essence, to raise the supply air temperature, the data centre's peak temperature should be minimised. It can be done through various means, including thermal aware workload scheduling and avoiding thermal imbalance in the data centre.

Application of ML: In one of the earlier studies, Google used ML for cooling system optimisation in their data centres

[21]. The study employed a neural network framework to model and predict the Power Usage Effectiveness (PUE), which is the ratio of the total building energy usage to the IT energy usage. It used historical data from servers and other cooling systems, such as server IT load, server temperature, cooling set points, outside temperature, and more. The study analysed the effect of PUE on various configurations and provided feedback for system administrators to fine-tune the configurations efficiently, such as the number of dry coolers running, water pump speed, and the number of process water pumps. With the abilities of ML models to capture the complex nonlinear behaviour between parameters of different subsystems, they have high potential in modelling of the CDC cooling systems and help administrators increase the energy efficiency by adjusting the knobs.

2) Thermal-aware IT Workload Management: Thermal aware workload management include many sub tasks such as workload scheduling, workload consolidation, and workload dispersion, among others. These tasks significantly affects the thermal bheaviours of a data centre. For instance, if the workload scheduling strategy results in peak temperature in the data centre, it generates a higher thermal load, thus increasing cooling costs. To address this, many researchers have proposed thermal-aware scheduling methods in Cloud data centres. Some solutions are proactive, which intends to avoid adverse temperature effects beforehand. In contrast, some scheduling policies follow reactive approaches. If a temperature violation is found, workloads are rescheduled to other nodes; however, the reactive scheduling method may result in higher QoS violations for applications due to rescheduling and migration. Mhedheb et al. [82] investigated load and thermal aware scheduling in Cloud that optimises temperature and load while scheduling tasks in data centres. Sun et al. [83] proposed thermal-aware scheduling of HPC jobs. They have used analytical models to estimate server temperature and model heat recirculation in the data centre. Proposed thermal-aware job assignment heuristics have shown increased performance with thermal balancing. Furthermore, authors in [84] have further extended thermal aware batch job scheduling across geo-distributed data centres.

Similarly, thermal-agnostic workload consolidation and dispersion triggers adverse temperature effects. Hence, balancing the workloads efficiently is necessary to achieve better efficiency. Consolidation is a widely used technique to optimise a computing system's energy consumption. However, aggressive consolidation leads to the creation of hotspots that further increases cooling cost. Hence, thermal-aware consolidation is necessary to balance the computing and cooling system energy consumption. A few studies have proposed solutions for this [85]–[87] to balance the temperature response due to workload placement during the workload consolidation. In contrast to to consolidation, the workload dispersion technique aims to spread out workloads evenly across the data centre's servers [88], preventing peak utilisation in normal conditions. Although it minimises peak temperature, it significantly increases the computing system energy due to resource under-utilisation. Hence, there should be a balance between consolidation and workload dispersion techniques to achieve cooling system efficiency.

Application of ML: Many of the existing works have employed machine-learning-based techniques in thermal-aware scheduling. Xiao et al. [89] presented a power and thermal-aware VM management framework based on machine learning, which relied on used Q-learning model to find optimal host configuration (power states) based on workload characteristics and cooling system's working state. The framework also enforced an efficient load-balancing policy to achieve a better trade-off between energy efficiency and performance. Similarly, many works have explored efficient distribution of application workload and also consolidation of VMs to increase resource utilisation and avoid thermal hotpots [90]–[92]. These works either develop temperature prediction model, aiding scheduling algorithms or develop controllers based on RL framework.

3) Thermal Modelling: Thermal modelling in data centres plays a vital role in resource management. Thermal modelling includes capturing thermal behaviour in a data centre and accurately estimating server temperature. Thermal models that predict accurately and fastly are useful aids in scheduling, configuring cooling systems and other resource management techniques. However, temperature prediction is a difficult problem. Server ambient temperature in a data centre depends on multiple factors, including CPU heat dissipation, inlet temperature and complex heat recirculation effects. There are mainly three types of thermal modelling techniques in data centres: (1) Computational Fluid Dynamic (CFD)-based models; (2) Analytical models; and (3) Predictive models.
The CFD models accurately capture the room layouts, and heat recirculation effects and accurately estimates temperature in the data centre [93]–[95]. However, they are computationally expensive, and even a single calibration requires models to be run for multiple days. Hence, they are incapable of using them for fast online resource management decisions. On the other hand, analytical models depends on modelling data centre and workloads based on mathematical frameworks [83], [96]. They represent cooling, computing and workload elements with formal mathematical models and build a framework to establish relationships between all elements [83]. Although they are fast in temperature estimation, their accuracy is compromised due to their rigid static models.

Application of ML: ML-based predictive models use actual measurement data from the data centre to predict the accurate temperature of the server. These data-driven models, once trained, are accurate and quickly deliver the results in runtime. Moreover, they can automatically model the physical layout, air conditioning and the heat generated by Cloud data centres. Unlike CFD, where each of these needs to be modelled explicitly, this is a huge benefit. To that end, Wang et al. [97] proposed a server temperature prediction model using the Artificial Neural Network (ANN) based ML technique. Results have shown that it can accurately predict the temperature in data centres. In addition, some studies have explored using

machine learning models to identify temperature distribution [98] and to predict server inlet temperature [99].

The drawback of the ML-based model is that the model is only applicable to the data centre which the data is collected from. This means data need to be collected for each data centre extensively. However, this is not a massive disadvantage as such data need to be collected to monitor the data centres' health.

C. Cooling Technologies for Thermal Management in Data Centres

When servers/IT equipment uses electricity for their operations, the electrical energy is transferred as heat. This heat will be drawn across the server cabinet by the rear-mounted server fans allowing heat to transfer from the server's components to the outside ambient environment. Many technologies are employed to take out this heat from the data centre environment and keep the data centre's operational temperature within its threshold. These cooling technologies can be broadly categorised into two categories, including air and liquid cooling technologies.

1) Air Cooling: Air cooling is a widely used data centre cooling technology due to its inexpensive and flexible design and operational conveniences. In rack-layout-based data centres, the dissipated heat from servers is extracted from the cooling system's environment. The **Computer Room Air Conditioning** (CRAC) is a cooling system responsible for monitoring and managing the temperature in the data centre [100]. The CRAC blows cold air through the perforated tiles under the racks of a data centre. The cold air passes from the bottom to the top of the rack taking out the dissipated heat from rack equipment sand this hot exhaust air is pushed to the intake of the CRAC units to the ceiling of the room, where it is taken out of the room. This allows the separation of hot exhaust air from cold inlet air. The CRAC unit then transfers the hot exhaust air via a coil to a fluid using refrigerant.

Many data centres also equip **Computer Room Air Handler (CRAH)**, where chilled water is used as fluid [100]. These fluids remove the heat from the data centre environment. The CRAC/CRAH continuously blow cold air using constant-speed fans, and this returns cold-air temperature, also called inlet temperature. It is configured to manage the dynamic thermal threshold in the data centre. It directly controls the cost of cooling in general. Lower the inlet temperature higher will be the cooling energy cost due to the increased energy required to transfer the lower temperature air from CRAC/CRAH. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) [80], a leading technical Committee in cooling system technology, recommends that the device inlet be between 18-27°C for the safe operation of the environment. The design goal of any data centre operators will be to provide an inlet temperature close to 27 °C to reduce the cooling cost. However, the safer operation threshold should be maintained while configuring this parameter. Many works have looked into optimising this parameter using different techniques by minimising the peak temperature [96] by balancing the workloads [92] and optimally configuring other parameters [101] of the cooling system.

Some modern systems also use **evaporative** [102] and **air side economisers/ free cooling** techniques [103]. In the evaporative technique, instead of fluid refrigerant, the hot air carried from the data centre is directly exposed to water. Water evaporates, taking out the heat from the hot air. Cooling towers are employed to dissipate the excess heat to the outside atmosphere. However, it doesn't require expensive CRAC or CRAH units but needs a large amount of water, a limiting factor in many data centre locations. On the other hand, air-side economisers or free cooling methods use outside free air for direct cooling instead of depending on the fluids to cool down the hot air extracted from CRAC/CRAH. This saves a huge amount of cooling costs. Nonetheless, these techniques vastly depend on the weather and geographical condition where the data centre is located, and thus they are used in limited computing infrastructures in practice.

2) Liquid Cooling: The recent advancement in data centre cooling technology has seen the adoption of liquid cooling as it is more efficient than air cooling, in general, [104]. The liquid cooling system also effectively avoids heat mix-up and heat re-circulation issues, which is a common problem in air cooling techniques.

Direct liquid cooling. In this system, liquid pipes are used to deliver liquid coolant directly to the heat sink present in the server's motherboards. The dissipated heat from the server is extracted to heat the chiller plant from these pipes, where the chilled water loop takes out the heat extracted from servers.

Immersion cooling. The computing system (servers and networking equipment is directly immersed in a non-conductive liquid. The liquid absorbs the heat and transfers it away from the components [105]. In some cases, equipment is arranged in isolated cabinets and immersed in tanks or cabinets are directly immersed in natural water habitats such as lakes/oceans. For instance, Microsoft has tested an underwater data centre with their project Natick [106], which allows them to operate the data centre in an energy-efficient manner by leveraging heat-exchange techniques with outside water. This technique is commonly used in submarines. This experimental project shows that immersion cooling is viable in large-scale computing systems with a group of servers sealed into large submarine cabinets.

Some other techniques have also been explored but are rarely used in large-scale settings, such as Dielectric fluid, where server components are coated with a non-conductive liquid. The heat is removed from the system by circulating liquid into direct contact with hot components, then through cool heat exchangers. Such methods are not widely adopted yet in practice. The common issue with rack-level liquid cooling is a lack of standardisation and specifications among multi-vendors. However, due to its energy efficiency compared to air cooling, it is expected that liquid cooling will become mainstream in future data centre cooling systems.

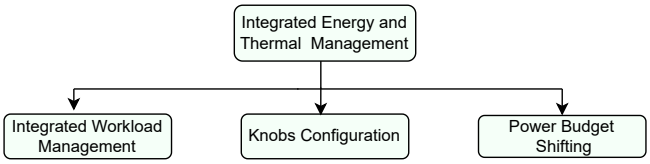

Fig. 6: Taxonomy of Integrated Energy and Thermal Management in Cloud Data Centres

V. Taxonomy of Integrated Energy and Thermal Management in CDCs

Traditionally cooling systems and computing systems are optimised individually. However, these two subsystems in the data centre are closely interdependent and optimising one system often have a opposite effect on others. Hence, the joint optimisation of two subsystems is beneficial, but it is challenging task that requires capturing complex dynamics of data centre workloads and physical environments. Fig. 6 shows a taxonomy of existing resource management solutions in integrated management of both computing and cooling energy.

Workload Management. A few studies have proposed solutions, including workload scheduling and cooling system optimisation as a multiobjective optimisation problem and accordingly configure different parameters to minimise energy consumption holistically [107], [108]. Other techniques include CRAC fan speed management by interplaying with IT load and its heat dissipation, configuring supply air temperature, and distributing the workload to minimise peak temperature, among many others.

Knobs Configurations. Wan et al. [109] studied holistic energy minimisation in data centres through a cross-layer optimisation framework for cooling and computing systems. This energy minimisation problem is formulated as a mixed-integer nonlinear programming problem. To solve this problem, the authors proposed a heuristic algorithm called JOINT, that dynamically configures parameters (such as server frequency, fan speed, and CRAC supply air temperature) based on workload demand and minimises computing and cooling system energy holistically.

Li et al. proposed [110] joint optimisation of computing and cooling systems for energy minimisation in data centres by modelling IT systems interactions (load distributions) and their corresponding thermal behaviour, i.e., heat transfer. The proposed analytical models for load distribution across rack servers minimise computing and cooling system energy, thereby configuring different knobs of two systems while ensuring the required throughput and resource constraints of workloads.

Power Budget Shifting. Power budget shifting is another important resource management technique in the Join optimisation of these two systems. Using available power to trade between two systems in runtime can increase energy efficiency and resource utlisaiton. PowerTrade [111] is a technique that trades off data centre computing systems' idle power and cooling power with each other to reduce total power. Over-provisioning is necessary for such conditions to accommodate extra workload and use excessive power obtained.

Application of ML: ML techniques have also been explored in the joint optimisation of computing and cooling systems. Recent advancements in RL have made it possible to learn different policies by interacting with the environments and learning from experience. RL techniques can be more adaptive and automatically understand the interdependence's of parameters. Ran et al. [112] used DRL and designed a hybrid action space that optimises the IT system and the airflow rate of the cooling system. Furthermore, the proposed control mechanism coordinates both the IT system's workload and cooling systems for energy efficiency. Similar techniques can be found in other studies [43], [113]. Careful design of state management, action, and rewards are important for applying RL techniques to data centres' holistic energy management.

VI. Future Research Directions

The sustainability in CDCs can be achieved by tackling some key issues that demand careful investigation and solutions. We need to fundamentally rethink how the data centres are currently managed, from hardware level optimisation to geo distributed data centre management. According to [11], if energy-aware approaches are implemented in CDCs, we can reduce total energy consumption in data centres up to 80% from the expected worst case scenario (Fig. 2). In the following, we identify key future research directions that should be pursued in this direction in order to reduce the energy footprints of CDCs and briefly explain them.

A. Standardisation and Tools for AI-centric RMS

One of the main barriers in adopting AI or ML solutions in data centre RMS is the lack of standardisation and tools. ML solutions need a lot of data. Currently, distributed systems, including Cloud systems, produce huge amounts of data from different computing layers. Standard methods and semantics are needed to collect, monitor, and interpret these data to accelerate the adoption of AI-centric models. Moreover, software tools and libraries need to be developed specifically for resource management systems, which will easily integrate policies into existing systems.

B. Hardware Software Co-design for ML-driven Resource Management

Computing servers and their components are tightly bound to operating systems that use simple rules to manage resources. This makes it hard to integrate new resource management policies that use ML to optimise hardware performance, because different vendors do not have common interfaces that can communicate with software. To solve this problem, we need a hardware-software co-design approach that allows us to develop and implement new resource management policies on hardware resources in an interoperable way.

C. Moving from "time-to-solution" to "Kw-to-solution"

Software development paradigms, platforms, and algorithms aim to enhance the execution speed of applications, but ignore their energy consumption. Hence, a paradigm shift is required from "time-to-solution" to "Kw-to-Solution" in software development and deployment. We also need new tools and programming constructs that enable software developers to assess and minimize the energy cost of their application logic, while preserving high speed. ML methods can offer valuable techniques for achieving this goal, such as learning energy-efficient code patterns, optimizing code performance, and adapting to different hardware configurations.

D. Resource Management in Emerging Cloud Workload Models using ML

Cloud computing is evolving from partially managed to fully managed services with application execution models like Serverless computing. Serverless computing lets us build applications with multiple stateless microservices. Cloud service providers handle the lifecycle of these microservices or stateless functions with guaranteed automatic scalability. This creates new challenges in pricing and managing thousands of stateless application services. ML-driven solutions, such as predicting user requests to cache "Hot" functions and reduce serverless function latency, or identifying resource interference among different user functions with classification methods, are some of the promising ways to address these challenges.

E. ML-Driven Holistic Resource Management

Cloud data centres consist of computing, networking, storage and cooling systems that are interdependent and crucial for ensuring service reliability. ML-driven resource management can detect these interdependencies and optimize the resources in a holistic manner to minimize energy consumption. A promising approach is to develop new algorithms and platforms that adjust parameters across different subsystems and balance tradeoffs.

F. Decarbonising Cloud Computing using ML

Cloud data centres are a major source of CO_2 emissions due to their dependence on fossil fuel-based energy sources. To decarbonise Cloud systems, many service providers are investing in renewable energy. However, the adoption of renewable energy sources is limited by their intermittent availability. Therefore, new solutions are needed to address the challenges of energy storage and workload management under uncertain energy supply. One promising direction is to use ML models to forecast the amount of renewable energy available at different Cloud data centre locations for a given time period. This prediction can enable the planning and execution of workloads in data centres that have more renewable energy, and thus reduce the reliance on fossil fuels.

G. Data-Driven Methods for Sustainable Multi-tier Computing Platforms

Multi-tier computing paradigms, such as Edge/Fog computing, have emerged to support IoT applications with distributed computations from the network edge to remote clouds. These paradigms pose new challenges for resource and application management, as they require low latency response and entail moving Cloud services from centralised locations to the network edge. Moreover, these paradigms involve more heterogeneous and energy-constrained environments than remote Clouds. Therefore, new solutions and approaches are needed for effective application and resource management under these conditions. ML methods can offer promising techniques for addressing these challenges, such as learning optimal resource allocation strategies, predicting workload patterns, and adapting to dynamic environments.

VII. Conclusions

Cloud computing platforms enable the development of highly connected resource-intensive applications, but they also require massive, heterogeneous, and complex data centres as their backbone infrastructure. Managing the energy and thermal aspects of such data centres is a challenging task, as the existing rule-based or heuristics solutions are not adequate to cope with the scale, heterogeneity, and dynamicity of the Cloud environment. Therefore, we need data-driven AI solutions that can leverage the data, learn from the environment, and make optimal resource management decisions. In this paper, we have explored leveraging AI-centric solutions for energy and thermal management in Cloud data centres. We have proposed a taxonomy for classifying different resource management techniques. We have also surveyed the state-of-the-art techniques and highlighted their strengths and limitations. Finally, we have suggested some promising future research directions

References

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, no. 6, pp. 599–616, 2009.

[2] A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, *et al.*, "Above the clouds: A berkeley view of cloud computing," *University of California, Berkeley, Rep. UCB/EECS*, vol. 28, no. 13, p. 2009, 2009.

[3] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[4] Amazon, " Amazon Web Services."

[5] Gartner, "Gartner forecasts worldwide public cloud revenue to grow 17 percent in 2020." https://www.gartner.com/en/newsroom/press-releases/2019-11-13-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-17-percent-in-2020, 2019. [Online; accessed 10-Jan-2021].

[6] Gartner, "Gartner forecasts worldwide public cloud end-user spending to grow 18 percent in 2021." https://www.gartner.com/en/newsroom/press-releases/2020-11-17-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-grow-18-percent-in-2021, 2020. [Online; accessed 10-Jan-2021].

[7] R. Buyya, S. Ilager, and P. Arroba, "Energy-efficiency and sustainability in new generation cloud computing: A vision and directions for integrated management of data centre resources and workloads," *Software: Practice and Experience*, vol. 54, no. 1, pp. 24–38, 2024.

[8] S. Maybury., "How much Energy does your Data Centre Use?.." https://www.metronode.com.au/energy_usage/, 2017. [Online; accessed 05-Jan-2021].

[9] A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, and W. Lintner, "United states data center energy usage report," 2016.

[10] J. Koomey, "Growth in data center electricity use 2005 to 2010," *A report by Analytical Press, completed at the request of The New York Times*, vol. 9, 2011.

[11] A. S. Andrae and T. Edler, "On global electricity usage of communication technology: trends to 2030," *Challenges*, vol. 6, no. 1, pp. 117–157, 2015.

[12] P. Johnson and T. Marker, "Data centre energy efficiency product profile," *Pitt & Sherry, report to equipment energy efficiency committee (E3) of The Australian Government Department of the Environment, Water, Heritage and the Arts (DEWHA)*, 2009.

[13] H. Viswanathan, E. K. Lee, and D. Pompili, "Self-organizing sensing infrastructure for autonomic management of green datacenters," *IEEE Network*, vol. 25, no. 4, pp. 34–40, 2011.

[14] D. Minoli, K. Sohraby, and B. Occhiogrosso, "Iot considerations, requirements, and architectures for smart buildings—energy optimization and next-generation building management systems," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 269–283, 2017.

[15] Q. Liu, Y. Ma, M. Alhussein, Y. Zhang, and L. Peng, "Green data center with iot sensing and cloud-assisted smart temperature control system," *Computer Networks*, vol. 101, pp. 104–112, 2016.

[16] S. Saha and A. Majumdar, "Data centre temperature monitoring with esp8266 based wireless sensor network and cloud based dashboard with real time alert system," in *2017 Devices for Integrated Circuit (DevIC)*, pp. 307–310, IEEE, 2017.

[17] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *arXiv preprint arXiv:1907.10597*, 2019.

[18] D. Amodei and D. Hernandez, "Ai and compute," 2018. https://blog.openai.com/ai-and-compute.

[19] D. Jeff, "ML for system, system for ML, keynote talk in Workshop on ML for Systems, NIPS," 2018.

[20] R. Bianchini, M. Fontoura, E. Cortez, A. Bonde, A. Muzio, A.-M. Constantin, T. Moscibroda, G. Magalhaes, G. Bablani, and M. Russinovich, "Toward ml-centric cloud platforms," *Communications of the ACM*, vol. 63, no. 2, pp. 50–59, 2020.

[21] J. Gao, "Machine learning applications for data center optimization," *Google White Paper*, 2014.

[22] W. Torell, K. Brown, and V. Avelar, "The unexpected impact of raising data center temperatures," *Write paper 221, Revision*, 2015.

[23] V. Venkatachalam and M. Franz, "Power reduction techniques for microprocessor systems," *ACM Computing Surveys (CSUR)*, vol. 37, no. 3, pp. 195–237, 2005.

[24] E.-Y. Chung, L. Benini, and G. De Micheli, "Dynamic power management using adaptive learning tree," in *Proceedings of the 1999 IEEE/ACM International Conference on Computer-Aided Design. Digest of Technical Papers (Cat. No. 99CH37051)*, pp. 274–279, IEEE, 1999.

[25] Y. Lu, D. Wu, B. He, X. Tang, J. Xu, and M. Guo, "Rank-aware dynamic migrations and adaptive demotions for dram power management," *IEEE Transactions on Computers*, vol. 65, no. 1, pp. 187–202, 2015.

[26] R. A. Bridges, N. Imam, and T. M. Mintz, "Understanding GPU power: A survey of profiling, modeling, and simulation methods," *ACM Computing Surveys*, vol. 49, no. 3, 2016.

[27] G. Von Laszewski, L. Wang, A. J. Younge, and X. He, "Power-aware scheduling of virtual machines in dvfs-enabled clusters," in *Proceedings of the 2009 IEEE International Conference on Cluster Computing and Workshops*, pp. 1–10, IEEE, 2009.

[28] P. Arroba, J. M. Moya, J. L. Ayala, and R. Buyya, "Dynamic voltage and frequency scaling-aware dynamic consolidation of virtual machines for energy efficient cloud data centers," *Concurrency and Computation: Practice and Experience*, vol. 29, no. 10, p. e4067, 2017.

[29] M. Safari and R. Khorsand, "Energy-aware scheduling algorithm for time-constrained workflow tasks in dvfs-enabled cloud environment," *Simulation Modelling Practice and Theory*, vol. 87, pp. 311–326, 2018.

[30] S. Wang, Z. Qian, J. Yuan, and I. You, "A dvfs based energy-efficient tasks scheduling in a data center," *IEEE Access*, vol. 5, pp. 13090–13102, 2017.

[31] J.-G. Park, N. Dutt, and S.-S. Lim, "Ml-gov: A machine learning enhanced integrated cpu-gpu dvfs governor for mobile gaming," in *Proceedings of the 15th IEEE/ACM Symposium on Embedded Systems for Real-Time Multimedia*, pp. 12–21, 2017.

[32] S. Ilager, R. Wankar, R. Kune, and R. Buyya, "Gpu paas computation model in aneka cloud computing environments," *Smart Data: State-of-the-Art Perspectives in Computing and Applications*, p. 19, 2019.

[33] X. Xu, W. Dou, X. Zhang, and J. Chen, "Enreal: An energy-aware resource allocation method for scientific workflow executions in cloud environment," *IEEE Transactions on Cloud Computing*, vol. 4, no. 2, pp. 166–179, 2015.

[34] H. Li, J. Li, W. Yao, S. Nazarian, X. Lin, and Y. Wang, "Fast and energy-aware resource provisioning and task scheduling for cloud systems," in *Proceedings of the 18th International Symposium on Quality Electronic Design (ISQED)*, pp. 174–179, IEEE, 2017.

[35] A. Beloglazov, J. Abawajy, and R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing," *Future Generation Computer Systems*, vol. 28, no. 5, pp. 755–768, 2012.

[36] M. Dabbagh, B. Hamdaoui, M. Guizani, and A. Rayes, "Energy-efficient resource allocation and provisioning framework for cloud data centers," *IEEE Transactions on Network and Service Management*, vol. 12, no. 3, pp. 377–391, 2015.

[37] H. Chen, X. Zhu, H. Guo, J. Zhu, X. Qin, and J. Wu, "Towards energy-efficient scheduling for real-time tasks under uncertain cloud computing environment," *Journal of Systems and Software*, vol. 99, pp. 20–35, 2015.

[38] Q. Huang, S. Su, J. Li, P. Xu, K. Shuang, and X. Huang, "Enhanced energy-efficient scheduling for parallel applications in cloud," in *Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012)*, pp. 781–786, IEEE, 2012.

[39] Y. Ding, X. Qin, L. Liu, and T. Wang, "Energy efficient scheduling of virtual machines in cloud with deadline constraint," *Future Generation Computer Systems*, vol. 50, pp. 62–74, 2015.

[40] R. N. Calheiros and R. Buyya, "Energy-efficient scheduling of urgent bag-of-tasks applications in clouds through dvfs," in *Proceedings of the 6th IEEE international conference on cloud computing technology and science*, pp. 342–349, IEEE, 2014.

[41] C. Ghribi, M. Hadji, and D. Zeghlache, "Energy efficient vm scheduling for cloud data centers: Exact allocation and migration algorithms," in *Proceedings of the 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing*, pp. 671–678, IEEE, 2013.

[42] J. L. Berral, Í. Goiri, R. Nou, F. Julià, J. Guitart, R. Gavaldà, and J. Torres, "Towards energy-aware scheduling in data centers using machine learning," in *Proceedings of the 1st International Conference on energy-Efficient Computing and Networking*, pp. 215–224, 2010.

[43] M. Cheng, J. Li, and S. Nazarian, "Drl-cloud: Deep reinforcement learning-based resource provisioning and task scheduling for cloud service providers," in *Proceedings of the 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 129–134, IEEE, 2018.

[44] A. Beloglazov, R. Buyya, Y. C. Lee, and A. Zomaya, "A taxonomy and survey of energy-efficient data centers and cloud computing systems," in *Advances in Computers*, vol. 82, pp. 47–111, Elsevier, 2011.

[45] S. F. Piraghaj, A. V. Dastjerdi, R. N. Calheiros, and R. Buyya, "A framework and algorithm for energy efficient container consolidation in cloud data centers," in *Proceedings of the 2015 IEEE International Conference on Data Science and Data Intensive Systems*, pp. 368–375, IEEE, 2015.

[46] M. H. Ferdaus, M. Murshed, R. N. Calheiros, and R. Buyya, "Virtual machine consolidation in cloud data centers using aco metaheuristic," in *Proceedings of the European conference on parallel processing*, pp. 306–317, Springer, 2014.

[47] N. T. Hieu, M. Di Francesco, and A. Ylä-Jääski, "Virtual machine consolidation with multiple usage prediction for energy-efficient cloud data centers," *IEEE Transactions on Services Computing*, vol. 13, no. 1, pp. 186–199, 2017.

[48] S.-Y. Hsieh, C.-S. Liu, R. Buyya, and A. Y. Zomaya, "Utilization-prediction-aware virtual machine consolidation approach for energy-efficient cloud data centers," *Journal of Parallel and Distributed Computing*, vol. 139, pp. 99–109, 2020.

[49] F. Farahnakian, P. Liljeberg, and J. Plosila, "Energy-efficient virtual machines consolidation in cloud data centers using reinforcement learning," in *2014 22nd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, pp. 500–507, IEEE, 2014.

[50] D. Basu, X. Wang, Y. Hong, H. Chen, and S. Bressan, "Learn-as-you-go with megh: Efficient live migration of virtual machines," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 8, pp. 1786–1801, 2019.

[51] A. A. Bhattacharya, D. Culler, A. Kansal, S. Govindan, and S. Sankar, "The need for speed and stability in data center power capping," *Sustainable Computing: Informatics and Systems*, vol. 3, no. 3, pp. 183–193, 2013.

[52] P. Petoumenos, L. Mukhanov, Z. Wang, H. Leather, and D. S. Nikolopoulos, "Power capping: What works, what does not," in *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 525–534, IEEE, 2015.

[53] R. Azimi, M. Badiei, X. Zhan, N. Li, and S. Reda, "Fast decentralized power capping for server clusters.," in *HPCA*, pp. 181–192, 2017.

[54] Q. Wu, Q. Deng, L. Ganesh, C.-H. Hsu, Y. Jin, S. Kumar, B. Li, J. Meza, and Y. J. Song, "Dynamo: Facebook's data center-wide power management system," *ACM SIGARCH Computer Architecture News*, vol. 44, no. 3, pp. 469–480, 2016.

[55] C. Lefurgy, X. Wang, and M. Ware, "Power capping: a prelude to power shifting," *Cluster Computing*, vol. 11, no. 2, pp. 183–195, 2008.

[56] A. Gandhi, M. Harchol-Balter, R. Das, and C. Lefurgy, "Optimal power allocation in server farms," *ACM SIGMETRICS Performance Evaluation Review*, vol. 37, no. 1, pp. 157–168, 2009.

[57] H. Chen, C. Hankendi, M. C. Caramanis, and A. K. Coskun, "Dynamic server power capping for enabling data center participation in power markets," in *Proceedings of the 2013 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 122–129, IEEE, 2013.

[58] A. G. Kumbhare, R. Azimi, I. Manousakis, A. Bonde, F. Frujeri, N. Mahalingam, P. A. Misra, S. A. Javadi, B. Schroeder, M. Fontoura, and R. Bianchini, "Prediction-Based power oversubscription in cloud platforms," in *Proceedings of the 2021 USENIX Annual Technical Conference (USENIX ATC 21)*, (Berkeley, CA, USA), pp. 473–487, USENIX Association, July 2021.

[59] H. Zhang and H. Hoffmann, "Maximizing performance under a power cap: A comparison of hardware, software, and hybrid techniques," *ACM SIGPLAN Notices*, vol. 51, no. 4, pp. 545–559, 2016.

[60] M. Xu and R. Buyya, "Managing renewable energy and carbon footprint in multi-cloud computing environments," *Journal of Parallel and Distributed Computing*, vol. 135, pp. 191–202, 2020.

[61] A. Khosravi, L. L. Andrew, and R. Buyya, "Dynamic vm placement method for minimizing energy and carbon cost in geographically distributed cloud data centers," *IEEE Transactions on Sustainable Computing*, vol. 2, no. 2, pp. 183–196, 2017.

[62] U. Mandal, M. F. Habib, S. Zhang, B. Mukherjee, and M. Tornatore, "Greening the cloud using renewable-energy-aware service migration," *IEEE network*, vol. 27, no. 6, pp. 36–43, 2013.

[63] Í. Goiri, K. Le, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenhadoop: leveraging green energy in data-processing frameworks," in *Proceedings of the 7th ACM european conference on Computer Systems*, pp. 57–70, 2012.

[64] Y. Zhang, Y. Wang, and X. Wang, "Greenware: Greening cloud-scale data centers to maximize the use of renewable energy," in *Proceedings of the ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and Open Distributed Processing*, pp. 143–164, Springer, 2011.

[65] J.-P. Lai, Y.-M. Chang, C.-H. Chen, and P.-F. Pai, "A survey of machine learning models in renewable energy predictions," *Applied Sciences*, vol. 10, no. 17, p. 5975, 2020.

[66] J. Gao, H. Wang, and H. Shen, "Smartly handling renewable energy instability in supporting a cloud datacenter," in *Proceedings of the 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 769–778, IEEE, 2020.

[67] L. Lin and A. A. Chien, "Adapting datacenter capacity for greener datacenters and grid," in *Proceedings of the 14th ACM International Conference on Future Energy Systems*, pp. 200–213, 2023.

[68] A. Radovanović, R. Koningstein, I. Schneider, B. Chen, A. Duarte, B. Roy, D. Xiao, M. Haridasan, P. Hung, N. Care, S. Talukdar, E. Mullen, K. Smith, M. Cottman, and W. Cirne, "Carbon-aware computing for datacenters," *IEEE Transactions on Power Systems*, vol. 38, no. 2, pp. 1270–1280, 2023.

[69] P. Wiesner, I. Behnke, D. Scheinert, K. Gontarska, and L. Thamsen, "Let's wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud," in *Proceedings of the 22nd International Middleware Conference*, Middleware '21, (New York, NY, USA), p. 260–272, Association for Computing Machinery, 2021.

[70] S. Pagani, P. S. Manoj, A. Jantsch, and J. Henkel, "Machine learning for power, energy, and thermal management on multicore processors: A survey," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 1, pp. 101–116, 2018.

[71] D. Shin, S. W. Chung, E.-Y. Chung, and N. Chang, "Energy-optimal dynamic thermal management: Computation and cooling power co-optimization," *IEEE Transactions on Industrial Informatics*, vol. 6, no. 3, pp. 340–351, 2010.

[72] R. Ayoub, K. Indukuri, and T. S. Rosing, "Temperature aware dynamic workload scheduling in multisocket cpu servers," *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, vol. 30, no. 9, pp. 1359–1372, 2011.

[73] H. F. Sheikh, I. Ahmad, Z. Wang, and S. Ranka, "An overview and classification of thermal-aware scheduling techniques for multi-core processing systems," *Sustainable Computing: Informatics and Systems*, vol. 2, no. 3, pp. 151–169, 2012.

[74] Z. Wang, C. Bash, N. Tolia, M. Marwah, X. Zhu, and P. Ranganathan, "Optimal fan speed control for thermal management of servers," in *Proceedings of the International Electronic Packaging Technical Conference and Exhibition*, vol. 43604, pp. 709–719, 2009.

[75] A. Iranfar, F. Terraneo, G. Csordas, M. Zapater, W. Fornaciari, and D. Atienza, "Dynamic thermal management with proactive fan speed control through reinforcement learning," in *2020 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 418–423, IEEE, 2020.

[76] W. Zhang, Y. Wen, Y. W. Wong, K. C. Toh, and C.-H. Chen, "Towards joint optimization over ict and cooling systems in data centre: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1596–1616, 2016.

[77] A. H. Khalaj and S. K. Halgamuge, "A review on efficient thermal management of air-and liquid-cooled data centers: From chip to the cooling system," *Applied energy*, vol. 205, pp. 1165–1188, 2017.

[78] C. Nadjahi, H. Louahlia, and S. Lemasson, "A review of thermal management and innovative cooling strategies for data center," *Sustainable Computing: Informatics and Systems*, vol. 19, pp. 14–28, 2018.

[79] M. Tian, *Energy Optimization by Fan Speed Control for Data Centers*. PhD thesis, The George Washington University, 2019.

[80] ASHRAE, "American society of heating, refrigerating and air-conditioning engineers," 2018. URL. http://tc0909.ashraetcs.org/.

[81] R. Zhou, Z. Wang, C. E. Bash, A. McReynolds, C. Hoover, R. Shih, N. Kumari, and R. K. Sharma, "A holistic and optimal approach for data center cooling management," in *Proceedings of the 2011 American Control Conference*, pp. 1346–1351, IEEE, 2011.

[82] Y. Mhedheb, F. Jrad, J. Tao, J. Zhao, J. Kołodziej, and A. Streit, "Load and thermal-aware vm scheduling on the cloud," in *Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing*, pp. 101–114, Springer, 2013.

[83] H. Sun, P. Stolf, and J.-M. Pierson, "Spatio-temporal thermal-aware scheduling for homogeneous high-performance computing datacenters," *Future Generation Computer Systems*, vol. 71, pp. 157–170, 2017.

[84] M. Polverini, A. Cianfrani, S. Ren, and A. V. Vasilakos, "Thermal-aware scheduling of batch jobs in geographically distributed data centers," *IEEE Transactions on cloud computing*, vol. 2, no. 1, pp. 71–84, 2013.

[85] E. K. Lee, H. Viswanathan, and D. Pompili, "Vmap: Proactive thermal-aware virtual machine allocation in hpc cloud datacenters," in *2012 19th International Conference on High Performance Computing*, pp. 1–10, IEEE, 2012.

[86] S. Ilager, K. Ramamohanarao, and R. Buyya, "Etas: Energy and thermal-aware dynamic virtual machine consolidation in cloud data center with proactive hotspot mitigation," *Concurrency and Computation: Practice and Experience*, vol. 0, no. 0, p. e5221, 2019.

[87] J. V. Wang, C.-T. Cheng, and C. K. Tse, "A thermal-aware vm consolidation mechanism with outage avoidance," *Software: Practice and Experience*, vol. 49, no. 5, pp. 906–920, 2019.

[88] H. Shamalizadeh, L. Almeida, S. Wan, P. Amaral, S. Fu, and S. Prabh, "Optimized thermal-aware workload distribution considering allocation constraints in data centers," in *2013 IEEE international conference on green computing and communications and ieee internet of things and IEEE cyber, physical and social computing*, pp. 208–214, IEEE, 2013.

[89] P. Xiao, Z. Ni, D. Liu, and Z. Hu, "A power and thermal-aware virtual machine management framework based on machine learning," *Cluster Computing*, vol. 24, pp. 2231–2248, 2021.

[90] S. Akbar, R. Li, M. Waqas, and A. Jan, "Server temperature prediction using deep neural networks to assist thermal-aware scheduling," *Sustainable Computing: Informatics and Systems*, vol. 36, p. 100809, 2022.

[91] S. S. Gill, S. Tuli, A. N. Toosi, F. Cuadrado, P. Garraghan, R. Bahsoon, H. Lutfiyya, R. Sakellariou, O. Rana, S. Dustdar, *et al.*, "Thermosim: Deep learning based framework for modeling and simulation of thermal-aware resource management for cloud computing environments," *Journal of Systems and Software*, vol. 166, p. 110596, 2020.

[92] S. Ilager, K. Ramamohanarao, and R. Buyya, "Thermal prediction for efficient energy management of clouds using machine learning," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 5, pp. 1044–1056, 2020.

[93] J. Choi, Y. Kim, A. Sivasubramaniam, J. Srebric, Q. Wang, and J. Lee, "A cfd-based tool for studying temperature in rack-mounted servers," *IEEE Transaction on Computers*, vol. 57, no. 8, pp. 1129–1142, 2008.

[94] R. Romadhon, M. Ali, A. M. Mahdzir, and Y. A. Abakr, "Optimization of cooling systems in data centre by computational fluid dynamics model and simulation," in *2009 Innovative Technologies in Intelligent Systems and Industrial Applications*, pp. 322–327, IEEE, 2009.

[95] A. Almoli, A. Thompson, N. Kapur, J. Summers, H. Thompson, and G. Hannah, "Computational fluid dynamic investigation of liquid rack cooling in data centres," *Applied energy*, vol. 89, no. 1, pp. 150–155, 2012.

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

[97] L. Wang, G. von Laszewski, F. Huang, J. Dayal, T. Frulani, and G. Fox, "Task scheduling with ann-based temperature prediction in a data center: a simulation-based study," *Engineering with Computers*, vol. 27, no. 4, pp. 381–391, 2011.

[98] Y. Tarutani, K. Hashimoto, G. Hasegawa, Y. Nakamura, T. Tamura, K. Matsuda, and M. Matsuoka, "Temperature distribution prediction in data centers for decreasing power consumption by machine learning," in *Proceedings of the 7th IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 635–642, IEEE, 2015.

[99] R. Lloyd and M. Rebow, "Data driven prediction model (ddpm) for server inlet temperature prediction in raised-floor data centers," in *Proceedings of the 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 716–725, IEEE, 2018.

[100] B. Fakhim, M. Behnia, S. Armfield, and N. Srinarayana, "Cooling solutions in an operational data centre: A case study," *Applied Thermal Engineering*, vol. 31, no. 14-15, pp. 2279–2291, 2011.

[101] E. K. Lee, H. Viswanathan, and D. Pompili, "Proactive thermal-aware resource management in virtualized hpc cloud datacenters," *IEEE Transactions on Cloud Computing*, vol. 5, no. 2, pp. 234–248, 2015.

[102] B. Porumb, P. Ungureşan, L. F. Tutunaru, A. Şerban, and M. Bălan, "A review of indirect evaporative cooling operating conditions and performances," *Energy Procedia*, vol. 85, pp. 452–460, 2016.

[103] H. Zhang, S. Shao, H. Xu, H. Zou, and C. Tian, "Free cooling of data centers: A review," *Renewable and Sustainable Energy Reviews*, vol. 35, pp. 171–182, 2014.

[104] T. Gao, S. Shao, Y. Cui, B. Espiritu, C. Ingalz, H. Tang, and A. Heydari, "A study of direct liquid cooling for high-density chips and accelerators," in *Proceedings of the 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pp. 565–573, IEEE, 2017.

[105] A. Capozzoli and G. Primiceri, "Cooling systems in data centers: state of art and emerging technologies," *Energy Procedia*, vol. 83, pp. 484–493, 2015.

[106] B. Cutler, S. Fowers, J. Kramer, and E. Peterson, "Dunking the data center," *IEEE Spectrum*, vol. 54, no. 3, pp. 26–31, 2017.

[107] W. Zhang, Y. Wen, Y. W. Wong, K. C. Toh, and C.-H. Chen, "Towards joint optimization over ict and cooling systems in data centre: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1596–1616, 2016.

[108] X. Li, P. Garraghan, X. JIANG, Z. Wu, and J. Xu, "Holistic Virtual Machine Scheduling in Cloud Datacenters towards Minimizing Total Energy," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 6, pp. 1–1, 2017.

[109] J. Wan, X. Gui, R. Zhang, and L. Fu, "Joint cooling and server control in data centers: A cross-layer framework for holistic energy minimization," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2461–2472, 2017.

[110] S. Li, H. Le, N. Pham, J. Heo, and T. Abdelzaher, "Joint optimization of computing and cooling energy: Analytic model and a machine room case study," in *Proceedings of the 2012 IEEE 32nd International Conference on Distributed Computing Systems*, pp. 396–405, IEEE, 2012.

[111] F. Ahmad and T. Vijaykumar, "Joint optimization of idle and cooling power in data centers while maintaining response time," *ACM Sigplan Notices*, vol. 45, no. 3, pp. 243–256, 2010.

[112] Y. Ran, H. Hu, X. Zhou, and Y. Wen, "Deepee: Joint optimization of job scheduling and cooling control for data center energy efficiency using deep reinforcement learning," in *Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 645–655, IEEE, 2019.

[113] A. Mirhoseini, H. Pham, Q. V. Le, B. Steiner, R. Larsen, Y. Zhou, N. Kumar, M. Norouzi, S. Bengio, and J. Dean, "Device placement optimization with reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 2430–2439, JMLR. org, 2017.