IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.DOI

A Review of Data Centers Energy Consumption And Reliability Modeling

KAZI MAIN UDDIN AHMED[1], (Student Member, IEEE), MATH H. J. BOLLEN[2], (Fellow, IEEE), AND MANUEL ALVAREZ.[3], (Member, IEEE)

[1]Electric Power Engineering, Luleå University of Technology, Forskargatan 1, 931 87 Skellefteå, Sweden.

Corresponding author: Kazi Main Uddin Ahmed (e-mail: kazi.main.uddin.ahmed@ltu.se).

This study is supported by the Swedish Energy Agency under Grant 43090-2, and in part by the Cloudberry Datacenters project, by Region Norrbotten, and by an industrial group.

ABSTRACT Enhancing the efficiency and the reliability of the data center are the technical challenges for maintaining the quality of services for the end-users in the data center operation. The energy consumption models of the data center components are pivotal for ensuring the optimal design of the internal facilities and limiting the energy consumption of the data center. The reliability modeling of the data center is also important since the end-user's satisfaction depends on the availability of the data center services. In this review, the state-of-the-art and the research gaps of data center energy consumption and reliability modeling are identified, which could be beneficial for future research on data center design, planning, and operation. The energy consumption models of the data center components in major load sections i.e., information technology (IT), internal power conditioning system (IPCS), and cooling load section are systematically reviewed and classified, which reveals the advantages and disadvantages of the models for different applications. Based on this analysis and related findings it is concluded that the availability of the model parameters and variables are more important than the accuracy, and the energy consumption models are often necessary for data center reliability studies. Additionally, the lack of research on the IPCS consumption modeling is identified, while the IPCS power losses could cause reliability issues and should be considered with importance for designing the data center. The absence of a review on data center reliability analysis is identified that leads this paper to review the data center reliability assessment aspects, which is needed for ensuring the adaptation of new technologies and equipment in the data center. The state-of-the-art of the reliability indices, reliability models, and methodologies are systematically reviewed in this paper for the first time, where the methodologies are divided into two groups i.e., analytical and simulation-based approaches. There is a lack of research on the data center cooling section reliability analysis and the data center components' failure data, which are identified as research gaps. In addition, the dependency of different load sections for reliability analysis of the data center is also included that shows the service reliability of the data center is impacted by the IPCS and the cooling section.

INDEX TERMS data center, data center design, planning and operation, energy consumption modeling, data center reliability, reliability modeling.

I. INTRODUCTION

A. BACKGROUND

With the development of cloud based services and applications, the commercial cloud service providers like Google, Facebook, or Amazon are now deploying massive geo-distributed data centers. According to a research conducted by the International Data Corporation (IDC), the global demand for the data transfer and digital services is expected to be doubled to 4.2 Zettabytes per year, equivalent to 42,000 Exabyte by 2022 [1]. The number of data centers is increasing globally to handle this rapidly growing data traffic, while the energy demand of the data centers is also increasing. According to [2], the US data centers handled about 300 million Terabyte of data that consumed around 8.3 billion kWh per year in 2016, hence 27.7 kWh per Terabyte with a carbon footprint of approximately 35 kg CO_2 per Terabyte of data.The Data Center Frontier has mentioned in a report that, the number of servers in data centers was increased

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

by 30% during 2010 – 2018 due to the growing demand of computational workloads [3]. With the growing number of servers, the number of computational instances including virtual machines running on the physical hardware was raised by 550%, the data traffic was climbed 11-fold, and the installed storage capacity was increased 26-fold during the same period [3]. Therefore, the global energy demand of the data centers grew from 194 TWh to 205 TWh during 2010 – 2018 [3].Additionally, the data centers will indirectly affect the CO_2 emission because of the growing energy demands, which has been projected up to 720 million tons by 2030 in [4]. [103]At present, the leading companies in the Information and Communication Technology (ICT) business are now building their new data centers in the high latitude areas in the Arctic region to avail the natural advantages including the renewable energy production facilities, the cold air and the appropriate humidity. Google has built a data center in Hamnia, Finland in 2011 to use the cold sea-water from the Bay of Finland and the onshore wind energy; while Facebook has moved to Sweden in 2013 and Ireland in 2016 for having natural advantages in the data center operation [4]. These companies are utilizing the natural advantages to reduce the energy consumption of the data centers, hence indirectly reducing their participation in the CO_2 emission. There are two major phases of data center innovation to cope with the challenges of energy efficiency. In the first phase, the data center operators have emphasized on improvement of efficiency of the Information Technology (IT) equipment and the data center cooling facilities during 2007 – 2014 [3]. During this time, the Nordic region has attracted significant investments for data centers for environmental benefits. For example, after Google and Facebook entered the region in 2009 and 2011, the Nordic countries have become a preferred site location by an increasing number of data center investors. A report by Business-Sweden estimates that the Nordics by 2025 could attract investments for data centers in the order of 2 – 4 billion Euro. This is based on the forecast of worldwide demand for data center services corresponding to the data center investments of the Nordic countries [5]. In the second phase, the large data center operators have focused on procuring renewable energy (i.e., wind, solar) to supply power for the data center operations instead of traditional power sources [3].

The data centers are opening new business opportunities while posing the following operational challenges:

- Increasing the energy efficiency of data centers to limit the energy consumption and CO_2 emission, hence reducing the operational cost of the data centers.
- Enhancing the service availability of the IT section, hence enhancing the overall reliability of the data center to satisfy the Service Level Agreements (SLA) with the clients of the data center.
- Making a strategical balance at the design stage to reduce the energy consumption and ensuring higher reliability of the data center.

The energy consumption and reliability models of the data center are needed to bring solutions for these two operational challenges in data centers. The energy consumption models could help to predict the consequences of the operational decisions, which results in more effective management and control over the system [6]. Furthermore, reliability modeling of the data center individual load sections and the reliability assessment of the data center as a whole are important to prevent unwanted interruptions in the services and to ensure the committed SLA [7]. In some cases, the reliability assessment model also demands the energy consumption models of the devices in load sections. As examples, the power losses of the Internal Power Conditioning System (IPCS) is taken into consideration to assess the overall service availability of the IT loads in [8], while the energy consumption models of the cooling section devices are used for cooling section's reliability assessment in [9]. In this regard, a suitable energy consumption model or modeling approach does not solely mean accuracy and precision of the model, while the energy consumption modeling approach of the data center often depends on the applications of the components' energy or power consumption models.

The purpose of this paper is to provide a review of the data center energy consumption modeling approaches and reliability modeling aspects that have been presented in the literature.

B. RESEARCH GAPS

The authors have reviewed 193 papers that are related to the data center energy consumption and the reliability modeling aspects. There exists a lack of review works in the literature regarding the data center reliability modeling, which is needed for further research to show the state-of-the-art of data center reliability analysis. In this paper, the authors have tried to fill the research gap by analyzing the reliability modeling aspects to show the current knowledge-base about data center reliability affecting factors, reliability indices, and reliability assessment methodologies.

Besides this, the energy and power consumption models of the data center loads are analyzed, and the advantages and disadvantages of the models to apply in research are explained, which are widely missing from the literature. Additionally, the power consumed by the devices in IPCS is also considered and analyzed as a data center load section like the IT and cooling load section, which is missing in previous review articles. As the trade-off between reducing the energy consumption and ensuring higher reliability of the data center is an operational challenge, which is not addressed properly in the literature. This paper gives recommendations to fill the research gaps by the future researchers for making a trade-off between the reliability and energy efficiency of data center.

C. OBJECTIVE AND APPROACH

The research interest in the energy-efficient and reliable operation of data centers has increased in last few decades, as shown in Figure 1. However, the number of the published

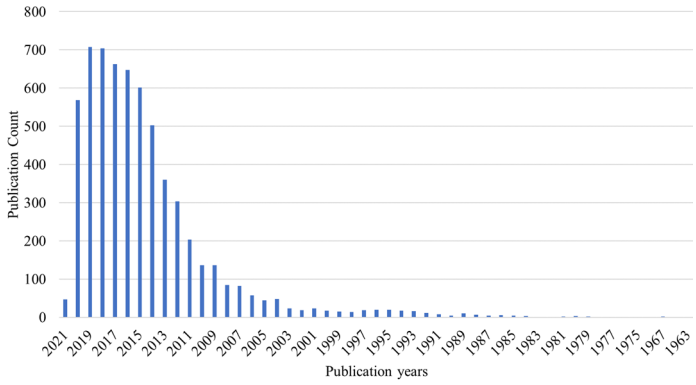
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

articles on data center reliability is lower than the number of articles on data center energy efficiency, which shows an urge to review the state-of-the-art of the data center reliability analysis. Moreover, the number of published articles on data center reliability analysis has reduced since 2016, as shown in Figure 1. Due to the lack of research in the data center reliability modeling the integration of new data center technologies could be impacted. The adaptation of the new technologies and equipment in the existing system of a data center depends on the reliability of the new technologies and equipment, which demands further research on it [10]. Apart from the reliability analysis, the energy efficiency analysis is also important for integrating the new technologies in data centers since most of the new technologies are coming with additional environmental challenges for the cooling load section [10]. Especially in the context of Green data center, which means the energy-efficient operation of the IT and the cooling load [10], the research on the data center reliability and energy consumption modeling should be emphasized. Therefore, the objective of this article is set to review the energy consumption models of the components in major load sections of the data center, and the data center reliability modeling including reliability indices, methodologies, and factors that affect the reliability of the data center. This review article could provide a potential starting point for further research on these topics.

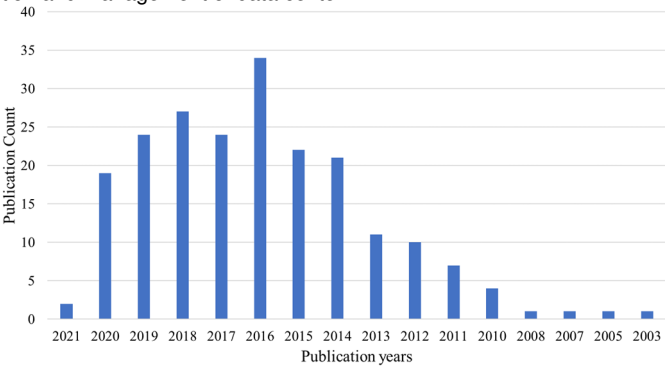
The authors intend to be as comprehensive as reasonable to select the published articles on related topics to review in this paper, however, it is not possible to guarantee that all the related papers are included. To obtain relevant papers, authors searched for the keywords "energy consumption and management" and "reliability assessment" in online databases like Google Scholar (http://scholar.google.com), Web of Science (https://apps.webofknowledge.com/), IEEExplore (http://ieeexplore.ieee.org), ACM Digital Library (http://dl.acm.org), Citeseer (http://citeseerx.ist.psu.edu), ScienceDirect (http://www.sciencedirect.com), SpringerLink (http://link.springer.com), and SCOPUS (http://www.scopus.com). Based on the searched results the research trends on the mentioned topics are shown in Figure 1, however, all the researched articles are not reviewed in this paper since the aim of this paper is to review the energy consumption models of the major components of the data center and the reliability modeling aspects of the data center. Therefore, the following keywords are used to filter the articles:

Power consumption modeling

- Data center loads, data center configurations
- Servers power consumption models (additive, base-active, regression, and server utilization based model)
- UPS, PDU, and PSU in data center application
- Power consumption model of UPS, PDU and PSU
- Data center cooling section
- Power consumption model of chiller, cooling tower, CRAC, and CRAH



(a) Publication trend searched with the keyword "Energy consumption and management of data center".



(b) Publication trend searched with the keyword "Reliability assessment of data center".

FIGURE 1: Web of Science indexed publication statistics on data center (Data collected in 27 February, 2021).

Reliability modeling

- Data center reliability analysis
- QoS and SLA of data center
- Tier classification of data center
- Service availability and service reliability
- Reliability modeling approach

The Systematic Literature Review (SLR) based methodological approach [11] is adopted in this paper with the above mentioned keywords. All the relevant attributes of the selected papers are used for constructing the knowledge base that is presented in this paper.

D. CONTRIBUTIONS & RECOMMENDATIONS

The contributions and the recommendations based on the review of energy consumption modeling of the data center are as follows:

- This paper has classified and summarized the published review articles based on their contributions for energy consumption modeling of the data center, while the absence of review articles on data center reliability assessment is also identified. Therefore, the data center reliability modeling aspects are comprehensively reviewed in this paper.
- The power and energy consumption models of the components and equipment in the major load sections of

the data center are reviewed in this paper. The proposed consumption models of the servers in IT load section are classified into four groups depending on the mathematical formulation of the models in the literature. The advantages, disadvantages, and applications of the server's power consumption models are also presented in this paper.

- The energy consumption models of the data center load sections are often used for analyzing the data center reliability, along with the aforementioned applications of the consumption models. The trade-off between the energy efficiency and the reliability of the data center is not addressed in research adequately that is found in the analysis of this paper.
  - -- Based on this analysis the recommendation for the future research on data center energy modeling would be choosing suitable energy consumption models of the equipment depending on the application. The accuracy of the models is often prioritize in research, however, it is found that the availability of the model parameters and variables are more important than the accuracy for research application. The energy consumption model parameters and variables that are easily accessible or measurable in laboratory facilities offer simplicity and ease in research applications.
  - -- More research should be conducted towards power losses and energy efficiency of the IPCS of the data center. There are research articles that present the load modeling for IT and cooling load section; this is not the case for the IPCS section, while the consumption of the IPCS section is found to be more than 10% of the total consumption.

The contributions and the recommendations based on the analysis of the data center reliability modeling and assessment techniques are as follows:

- This paper reviews the reliability modeling aspects related to the data center. The reliability indices and metrics for IT, IPCS, and the cooling load sections are analyzed including the reliability modeling methodologies. The reliability modeling methodologies are classified into two groups (i.e., analytical and simulation-based) depending on the modeling approaches. This research identifies the state-of-the-art of data center reliability modeling techniques that are studied so far, which could be a starting tool for future researchers.
- The need to have a standard code for data center operation along with the tier classification is identified in this paper since the failure and the degraded mode of a data center can impact the reliability differently.
  - -- The recommendation is to focus on the data center reliability study considering new equipment and topologies with new technologies. The new technologies are putting more stress on the load sections as explained in [10]. The lack of research on data center reliability aspects could hamper the development growth of the individual load section and also the development of the data center industry.
  - -- The lack of research on the data center's cooling load reliability is addressed; thus it is recommended to give more research focus on the cooling section reliability assessment.
  - -- The availability of data center component's statistical failure data is important for reliability studies at different levels of data centers. Thus, it is recommended to the data center owner/operators to publish the statistical failure data of data center components to ensure the adequacy of resources for further research.

E. ORGANIZATION

The paper is structured as follows: The contributions and remarks of the published related review papers are explained in Section II. Section III analyzes the energy consumption models of the data center's major load sections. Section IV discusses the reliability modeling aspects of the data center. The limitations and the future works are explained in Section V. Finally, Section VI concludes the article with recommendations and discussions based on the analysis.

II. RELATED REVIEW ARTICLES

According to the Web of Science at least 56 review articles have been published between 2005 – 2020, where the articles have presented the overview of the data center load section's energy and power consumption models and the application of the models in the data center. The articles are searched with the keywords "review OR overview OR survey data center energy consumption model" in the database. The published articles are classified into two categories: 1) the energy efficiency techniques at component-level to data center system-level, and 2) the energy management techniques. The energy management techniques also include the thermal environment design and management, air flow control including free cooling, thermal metrics, and thermal parameter optimization. Moreover, the researchers' interest in these articles is also increasing, which is depicted by the increasing number of citations of the articles. The review articles are analyzed based on the subjects of review and the number of citations, as shown in Table 1 and Table 2.

The review articles addressing "data center reliability or availability modeling" were not found. However, the research interests on data center reliability modeling have been observed by the increasing number of published articles that address various aspects of the data center reliability, as shown in Figure 1b, which also quests about a SLR considering the data center reliability modeling for future researchers.

A taxonomy based on the overview of the energy consumption and reliability modeling of the data center is shown in Figure 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

TABLE 1: Summary of review articles based on major contributions.

Subject of review	References
Airflow control, distribution, and management in data center	[12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]
Numerical load modeling of data center components	[6], [14], [23], [24], [25], [26], [27], [28], [29]
Proposed energy consumption models and validation with experimental results	[6], [25]
Methodologies to evaluate the data center performance, including performance metrics	[13], [16], [23], [30], [31], [28]
Dynamic behavior of data center load sections	[6]
Operational cost analysis	[30], [32], [33]
Power saving techniques including efficient operation of data center	[24], [26], [31], [34], [35], [36], [37], [38]
Computational workloads, data traffic management, and data traffic control for energy efficiency	[39], [40], [41], [42], [43]
Data center resources scheduling and management	[38], [40], [44], [45], [46], [47], [48], [49]
Data center architecture evaluation	[23], [28], [37], [41]

TABLE 2: Number of citations of the review articles.

Article	[28]	[37]	[42]	[18]	[41]	[16]	[19]	[47]	[32]	[33]	[20]	[24] [50]	[23]	[48]	[22] [43]
Nr. of Citations	261	116	114	91	81	62	30	25	13	12	8	7	6	3	2
Publishing Year	2016	2014	2016	2015	2014	2017	2015	2016	2019	2020	2015	2019 2011	2020	2017	2014 2016

III. REVIEW OF DATA CENTER LOAD MODELING

Data center accommodates ICT equipment, which provides data storage, data processing, and data transport services [51]. Data centers typically have three major load sections: IT loads, cooling and environmental control equipment, and internal power conditioning system i.e., Uninturrupted Power Supply (UPS), Power Distribution Unit (PDU), and Power Supply Unit (PSU), including security and office supports, as shown in Figure 3. The IT load section contains servers, storage, local cooling fans, network switches, etc. The data center also needs a power conditioning system with cooling and environmental control to maintain the adequate power quality and the required temperature for the IT loads [28], [52], [53]. The IT load section of the data center is needed to be environmentally controlled since it houses devices like servers and network switches that generate a considerable amount of heat. The IT devices are highly sensitive to temperature and humidity fluctuations, so a data center must keep restricted environmental conditions for assuring the reliable operation of its equipment [25]. Besides the IT and cooling load sections, the power conditioning section is another important part of the data center that also consumes power [8], [52]. The amount of power consumed by the load sections depends on the design of the data center and the efficiency of the equipment. The largest power consuming section in a typical data center is the IT load section including IT equipment (45%) and the network equipment (5%), while the cooling loads (38%) rank second in the power consumption hierarchy, as shown in Figure 4 [24], [52]. Besides these two load sections, the power conditioning devices in the IPCS consume 8% of the total power of the data center, which has not been studied deeply in the existing literature. However, consideration of every possible power consumption is needed to properly model the power consumption of the entire data center because a model is a formal abstraction of a real system [54]. Regarding the power consumption of the load sections in a data center, the models can be represented as equations, graphical models, rules, decision trees, sets of representative examples, neural networks, etc [28]. The following are the main applications of the power consumption models for the data center.

- **Design of the power supply system of a data center:** The power consumption models of the load sections are necessary for the initial design stage of the IPCS of energy intensive industries like the data centers. It is not worth building a IPCS without prior knowledge of energy demand load sections and the power losses of the system [55]. A simulation tool is proposed in [56] that evaluates the Power Usage Efficiency (PUE) and other energy usage efficiency factors of data centers, which is applied in the Data Center Efficiency Building Blocks project to optimize the energy consumption of the data center considering the maximum loads in the data center, as explained in [56]. The power consumption models of the load sections could be a useful tool to design the internal power supply infrastructure of the data center.
- **Forecasting the energy consumption trends and enhancing the energy efficiency of data centers:**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

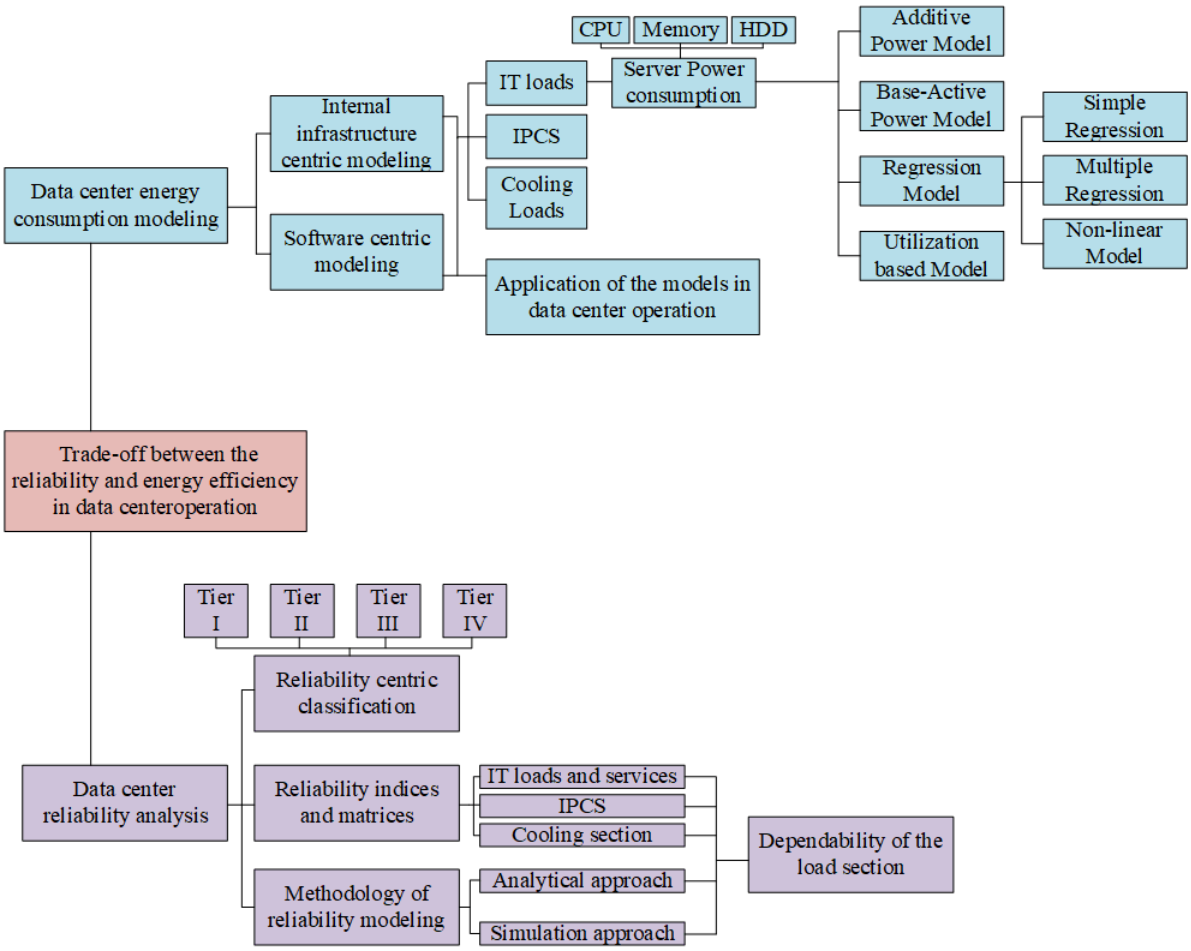


FIGURE 2: Taxonomy of energy consumption and reliability modeling of the data center.

Understanding the power consumption trend of the data center load sections is important for maximizing the energy efficiency. In data center operation, the real-time power measurements can not help to take decisions and provide the solutions, thus the predicted power consumption of the load sections is needed alongside [57]. The power consumption models of the data center components are used to predict the power consumption of the load sections in [58]. The forecasted power consumption trends of the load section helps in data center operation to optimize the overall consumption of the data center [59].

- **Power consumption optimization:**
  Different power consumption optimization models have been applied in data center using the power consumption models of the data center load sections to ensure the energy efficiency and cost effective data center operation. In [10], [60] the power consumption models of the load sections are used for optimizing the power consumption of the data center.

Modeling the exact power consumption behavior of a data center, either at the system level or at the individual component level, is not straightforward. The power consumption of a data center depends on multiple factors like the hardware specifications and internal infrastructure, computational workloads, type of applications of the data center, the cooling requirements, etc., which cannot be measured easily [10], [33]. Furthermore, the power consumption of the hardware in the IT load section, the cooling section, and the power conditioning infrastructure of the data center are all closely coupled [61]. The development of the component level power consumption models helps in different activities such as new equipment procurement, system capacity planning, resource expansion, etc. The power consumption models of different load sections are described in the following part of this section.

A. IT LOAD MODELS

Some of the discussed components in IT load section may appear at different other levels of the data center hierarchy, however, all of them are specifically attributed to IT loads of a data center. Traditionally the servers are the main computational resource in the IT load section. Other devices like memory, storage, network devices, local cooling fans, and

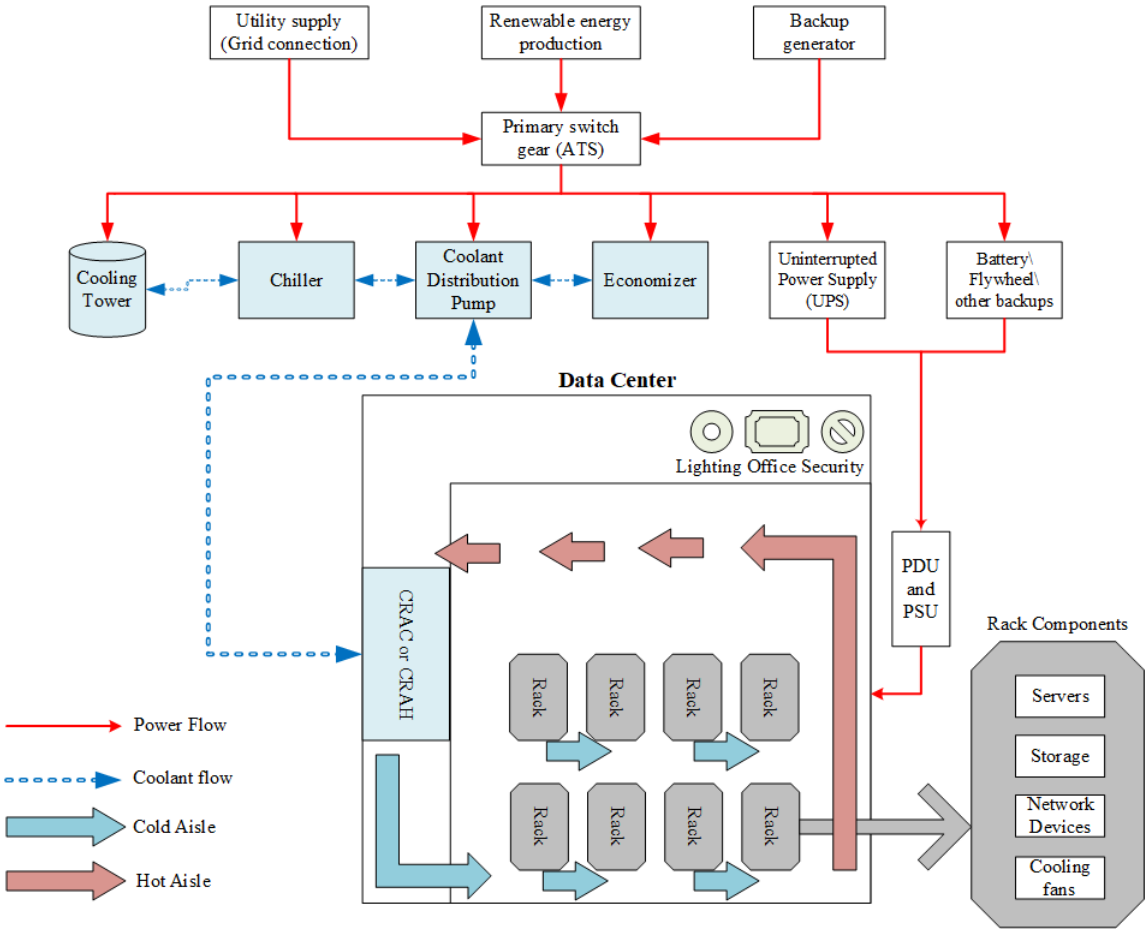


FIGURE 3: Internal structure of a typical data center.

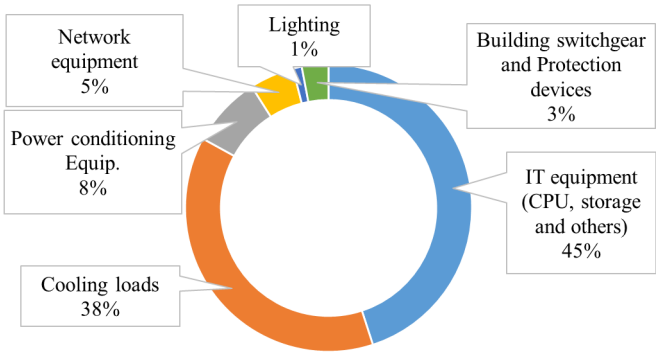


FIGURE 4: Analysis of power consumption proportionality in data center. [8], [24]

server power supplies are also considered as IT load in the literature. The most power consuming components in the IT load section are the servers [39]–[41]. The percentage of power consumption by the components of the servers is shown in Figure 5, [23], [24]. The Central Processing Unit (CPU) is the largest contributor to the total server power consumption, followed by peripheral slots (including network card slot and Input and Output devices (I/O) devices), conduction losses, memory, motherboard, disk/storage, and cooling fan. Therefore, the energy usage or the power consumption models of the server that has been presented in the literature are emphasized in this paper.

Server Consumption Model

The proposed power consumption models of the servers are classified into four groups based on the characteristics of the proposed power and energy consumption models, namely additive model, baseline-active model, regression model, and utilization-based model.

1) Additive Power Models

The power consumption models of the server that are proposed as a summation of the server components' power consumption belong to this group, as summarized in Table 3. The most simple server power consumption model was proposed considering the power consumption of the CPU and memory unit in [62]. Later, other additive models are proposed considering additional components in the equation of the server power or energy consumption model, as shown in Table 3. Most of the proposed models tried to mimic the power consumption of the main-board or motherboard as the power consumption of the servers like in [62]–[64], while

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

the consumption of motherboard is addressed separately in [65]. The power consumption of the motherboards can be considered as the conduction loss of the server, as shown in Figure 5.

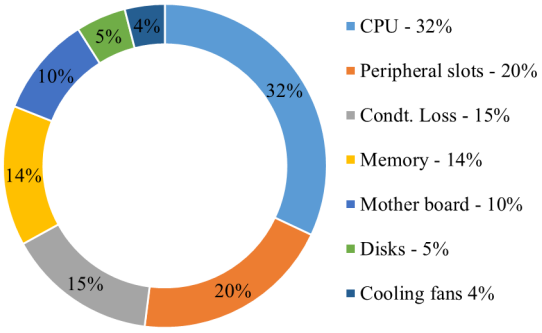


FIGURE 5: Component-wise energy consumption of a server. [23], [24]

A further extension of the additive server power consumption model is presented in [66], where the overall power consumption of the server is proposed with a base level consumption, P_{base}, as shown in (1). P_{base} accounts for the un-addressable power losses including the idle power consumption of the server.

$$P_{server} = P_{base} + P_{CPU} + P_{disk} + P_{net} + P_{mem} \tag{1}$$

The power modeling approach as shown in (1), can be further expanded considering the fact that the energy consumption can be calculated by multiplying the average power by the execution time [64]:

$$E_{total} = P_{comp} \cdot T_{comp} + P_{NIC} \cdot T_{comm} + P_{net_{dev}} \cdot T_{net_{dev}} \tag{2}$$

A different version of (1) is obtained by considering levels of resource utilization by the key components of a server [67]:

$$P_t = C_{cpu_n} \cdot u_{cpu_t} + C_{mem} \cdot u_{mem_t} + C_{disk} \cdot u_{disk_t} + C_{NIC} \cdot u_{NIC_t} \tag{3}$$

A similar energy consumption model of the entire server is described by Lewis et al. in [68]:

$$E_{system} = A_0 \cdot (E_{proc} + E_{mem}) + A_1 \cdot E_{em} + A_2 \cdot E_{board} + A_3 \cdot E_{hdd} \tag{4}$$

where, A_0, A_1, A_2, and A_3 are constants that are obtained via linear regression analysis and remain the same for a specific server architecture. The terms E_{proc}, E_{mem}, E_{em}, E_{board}, and E_{hdd} represent the total energy consumed by the processor, energy consumed by the DDR and SDRAM chips, energy consumed by the electromechanical components in the server blade, energy consumed by the peripherals that support the operation onboard, and energy consumed by the hard disk drive. The close relation between CPU and memory energy consumption is attributed by assigning the same constant A_0 for both CPU and memory.

2) Baseline - Active (BA) Power Model

In data centers, the servers do not always remain in the active state, as servers can be also switched to the idle mode. Therefore, the power consumption of the server can be divided into two parts, i.e., (1) Baseline power (P_{base}), and (2) Active power (P_{active}). The idle power consumption of the server also includes the power consumption of the fans, CPU, memory, I/O, and other motherboard components in their idle state, denoted by P_{base}. It is often considered as a fixed value [73], [74]. P_{active} is the power consumption of the server depending on the computational workloads, hence, on the server resource utilization (i.e., CPU, memory, I/O, etc). Therefore, the power consumption model can be expressed as the sum of the baseline power and active power, as given in (5). Similar server power consumption models related to the Base Active (BA) modeling approach are presented in Table 4.

$$P_{BA} = P_{base} + P_{active} + P_{\Delta} \tag{5}$$

where, P_{Δ} is the correction factor of the server power consumption model, which can be either a fixed value or an expression.

The active state power consumption of the server, P_{active} of the BA models can be expressed as a function of server utilization, coolant pump power consumption, Virtual Machine (VM) utilization factor, etc., as depicted in Table 4.

3) Regression Models

The regression model of the server power consumption considers the correlation between the power consumption and performance counters of the functional units of the servers i.e., CPU, memory, storage, etc. The regression models capture the fixed or idle power consumption and the dynamic power consumption with changing activity across the functional units of the servers. Therefore, the regression based server power consumption models are also known as 'Power Law models', which has become popular in data center application during $2010-2014$. The regression models are mostly adopted in research because of the simplicity and interpretability of the models, however, these models are not suitable to track the server power consumption in cloud interfaces since the server workloads fluctuate frequently [77]. The accuracy of the regression models are analyzed in [78], where it is mentioned that the regression models can predict the dynamic power usage well with the error below 5%. However, the error can be around 1% for non-linear models depending on the usage case [23]. In this paper, the regression models of the servers are classified into three groups (i.e., simple regression model, multi regression model, and non-linear model).

- Simple regression model
  The correlation between the power consumption and the performance counters that captured the activity of the CPU was first proposed in [79], while the mathematical model was first presented in [78], as given in (6). Addi-

TABLE 3: Summary of the proposed additive server power models found in the literature.

Equation	Limitations
$E_{server}(A) = E_{CPU}(A) + E_{mem}(A)$ [62]	The proposed model is specific to a predefined workload A, which is not feasible to analyze on a large scale setup with thousands of servers [63], [69]. However, this equation is suitable for bench-marking the prototypes, and for experimental result analysis [62].
$E_{server} = E_{CPU} + E_{mem} + E_{I/O}$ [63]	Current platforms with the compact structure of motherboards do not allow measuring the power consumption of individual component of servers separately [64].
$E_{server} = E_{CPU} + E_{mem} + E_{disk} + E_{NIC}$ [64], [70]	The proposed model considers more server components but still neglects at least two components i.e., the fans and motherboard consumption, which is criticized in [65], [71].
$P_{server} = \sum_1^i P_{mb_i} + \sum_1^j P_{Fan_j} + \sum_1^k P_{PSU_k}$ [72]	The PSU power consumption is added with the server consumption, which is not the case in practice. The PSUs are Apart of the IPCS. A similar modeling approach is used with modification in [52].

TABLE 4: Summary of the Base - Active models found in the literature.

Equation	Limitations
$P_{server} = P_{base} + \sum_1^M P_i^{VM}$ $P_i^{VM} = \frac{P_{base}}{M} + W_i \sum k_j \times U_j$ [74]	The VMs cannot be connected by hardware power meters, their actual power consumption P_M^{VM} are calculated as a function of server utilization.
$P_{server} = P_{fix} + P_{var}$ [75]	The cooling section power consumption is added with the server power and the model is proposed and verified for cloud systems.
$P_{server} = P_{IT} + \sum_1^n \cdot nP_{Pump} + \sum_1^m \cdot mP_{Fan}$ [76]	The proposed model is strictly valid for servers with liquid cooling.
$P_{server} = P_{IT} + P_{fan}$ [53]	P_{IT} and P_{fan} are further derived from a regression model based on experimental values.

tionally, the power consumption model presented in [78] was also validated by the experimental results.

$$P_{server} = P_{idle} + (P_{active} - P_{idle}) \cdot u \tag{6}$$

A similar model for cloud based system with VMs is presented in [80], which has the scope to use different independent variables for different application scenarios:

$$P_x = H_{idle} + (H_{active} - H_{idle}) \cdot \frac{U_x^{uti}}{\sum_{y=1}^{U_{count}} U_y^{uti}} \tag{7}$$

Similar simple regression models presented in different articles are summarized in Table 5.

- Multiple regression model

The simple regression models that are shown in (5)-(6) are based on CPU utilization, but addressed from different points of view. These power consumption models can provide reasonable accuracy for CPU-intensive workloads, however, cannot show the change in power consumption of servers caused by I/O and memory-intensive applications [73].

The server power consumption model as a function of utilization of the CPU, memory, disk, and network devices is presented in [84], as shown in (8). It assumes that subsystems such as CPU, disk, and I/O ports show a linear power consumption concerning their individual utilization, as discussed in [85].

$$\begin{aligned} P_{server} =& 14.45 + 0.236 \cdot u_{cpu} - (4.47 \times 10^{-8}) \cdot u_{mem} \\ &+ 0.00281 \cdot u_{disk} + (3.1 \times 10^{-8}) \cdot u_{net} \end{aligned} \tag{8}$$

A classified piecewise linear regression model is presented in [86] to achieve a more accurate power prediction, as shown in (9). It is noteworthy that n_{VM} in (9) is the number of VMs running on a server, which is assumed to homogeneous in configuration thus the weights α, β, γ and e for each VM are the same. The proposed model considers the components of the server to be connected as building blocks of the server, which is valid for blade servers. It also assumes that subsystems show a linear power consumption concerning their

TABLE 5: Summary of the simple regression models found in the literature.

Equation	Limitations
$P_{s[t1,t2]} = P_{idle} + \frac{Conn_{s[t1,t2]}}{Conn_{Max_s}} \cdot (P_{max_s} - P_{idle})$ [81]	The power consumption model of the server is proposed for the Content Distribution Network (CDN)s with different configurations, however, validating the proposed model is difficult with the real time applications of the CDN. Moreover, the application of this model is limited for general use-cases as it needs the number of connections $Conn_{s[t_1,t_2]}$ between time t_1 to t_2, which is not usual to avail in general use-cases.
$P_n = P_n^0 + (P_{max} - P_n^0) \cdot \frac{r_n}{r_n^*}$ [60]	The proposed model is developed based on a ratio of the goodput (output of an assigned task) r_n and the maximum supportable goodput r_n^*, which is further validated for virtualized servers in [60].
$P_\lambda = \frac{\lambda}{\mu} \cdot (P_{CPU} + P_{other}) + (1 - \frac{\lambda}{\mu}) \cdot P_{idle}$ [82]	The ratio $(\frac{\lambda}{\mu})$ defines the utilization of the server, therefore it is the same as given (6).
$P(u) = k \cdot P_{max} + (1-k) \cdot P_{max} \cdot u$ [83]	k is the fraction of power consumed by the idle server, which is not always available for measurement since it depends in the CPU utilization u, as given in (6).

individual utilization, as shown in (9). In this contrast, Kansal et al. proposed further detailed model of server power consumption in [87], considering CPU utilization, the number of missing Last Level Cache (LLC), and the number of bytes read and written, as shown in (10). These two consumption models are basically the same, except the additional term N_{LLCM}, as it is depicted by the comparison of (9) and (10). The term N_{LLCM} represents the number of the missing LLC during T, and α_{mem} and γ_{mem} are the linear model parameters. A more generalized power consumption model is presented in [88] based on the server's components performance counters (i.e., CPU cycles per second, references to the cache per second, cache misses per second), as given in (11). Later, the power consumption model in (11) is further extended by Witkowski et al. [89] by including the CPU temperature in the model.

$$P_{server} = \alpha \cdot \sum_{k=1}^{n} U_{CPU}(k) + \beta \cdot \sum_{k=1}^{n} U_{mem}(k) + \gamma \cdot \sum_{k=1}^{n} U_{I/O}(k) + e \cdot n_{VM} + P_{const} \quad (9)$$

$$E_{server} = \alpha_{CPU} \cdot u_{CPU}(p) + \gamma_{CPU} + \alpha_{mem} \cdot N_{LLCM} + \gamma_{mem} + \alpha_{io} \cdot b_{io} + \gamma_{disk} + E_{static} \quad (10)$$

$$P = P_0 + \sum_{i=1}^{I} \alpha_i \cdot Y_i + \sum_{i=1}^{J} \beta_j \sum_{l=1}^{L} X_{jl} \quad (11)$$

where the power consumption of a server by a combination of variables Y_i, $i = 1,...,I$, and X_{jl}, $j = 1,...,J$ describing individual processes l, $l = 1,...,L$. The power consumption of a server with no load is denoted by P_0 (the intercept), and the respective coefficients of the regression model are α_i and β_j.

The ambient temperature, CPU die temperature, memory and hard disk consumption, including the energy consumed by the electro-mechanical components are added to the regression model in [90], as shown in (12). These models can predict the energy consumption precisely as long as the trend of workload does not change.

$$E_{server} = \alpha_0 \cdot (E_{proc} + E_{mem}) + \alpha_1 \cdot E_{em} + \alpha_2 \cdot E_{board} + \alpha_3 \cdot E_{hdd} \quad (12)$$

- Non-Linear Models

  A non-linear model is proposed in [78] that includes a calibration parameter r, which minimizes the square error, as shown in (13). The square error needs to be obtained experimentally since it depends on the type of the server. This same model is also presented in [83], [91]

$$P_u = (P_{max} - P_{idle})(2u - u^r) + P_{idle} \quad (13)$$

  where r is a calibration parameter that minimizes the square error which needs to be obtained experimentally. The power model in (13) performs better than the regression models to project the power consumption of the servers [78], however, it needs to determine the calibration parameter r which is a disadvantage associated with the model. Meanwhile, Zhang et al. in [58] has used high-degree polynomial models to fit the server power consumption, finding that the cubic polynomial

model as in (14c) is the best choice compared to (14a) and (14b). Similarly, the relationship between power consumption and the second order polynomial of server utilization is provided in [92].

$$P_{total} = a + b \times R_{CPU} \tag{14a}$$

$$P_{total} = a + b \times R_{CPU} + c \times R_{CPU}^2 \tag{14b}$$

$$P_{total} = a + b \times R_{CPU} + c \times R_{CPU}^2 + d \times R_{CPU}^3 \tag{14c}$$

where R is the resource utilization, a, b, c, and d are the constants of the polynomial fit.

4) Utilization-Based Power Model

Most of the system utilization-based power models leverage CPU utilization as their metric of choice in modeling the entire server's power consumption since CPU is the most power consuming component in the server, as shown in Figure 5. One of the earliest CPU utilization-based server power models has appeared in [93], as shown in (15), which is an extension of the basic digital circuit power model, given in (16). The P_{dyn} is the dynamic power consumption of any circuit caused by capacitor switching, where A denotes the switching activity (i.e., Number of switches per clock cycle), C as the physical capacitance, V as the supply voltage, and f as the clock frequency. Different techniques can be applied for scaling the supply voltage and frequency in a larger range, as shown in (15).

$$P(f) = c_0 + c_1 \cdot f^3 \tag{15}$$

$$P_{dyn} = ACV^2 f \tag{16}$$

It is important mentioning the voltage is proportional to the frequency f as $V = (constant) \times f$ [93]. The constant c_0 includes the power consumption of all components except for the idle power consumption of the CPU in (15). The term $c_1 = AC(\frac{V}{f})^2$ are obtained from (16) where A and C is the switching activity (i.e., number of switches per clock cycle) and the physical capacitance, respectively.

Further in 2007, another notable CPU utilization-based power model is presented in [78] which has influenced recent data center power consumption modeling research significantly, as given in (17). This power consumption model of the server can track the dynamic power usage with a greater accuracy at the PDU level [94], [95]. This power consumption model of the server also fits into the catalog of simple regression models because of the mathematical formulation, as shown in (6).

$$P_{server} = (P_{max} - P_{idle}) \cdot u + P_{idle} \tag{17}$$

This model assumes that the server power consumption and the CPU utilization have a linear relationship. Studies have used this empirical model as the representation of the server's total power consumption in [58], [96]. However, certain researches define a different utilization metric for the power consumption model of the server. The power model defines the utilization as the percentage between the actual number of connections made to a server against the maximum number of connections allowed on the server in [81], which is used for a specific use-case to model the power consumption of a content delivery network server. A non-linear server power model based on CPU utilization is proposed in [83], [91], as shown before in (13).

Importance of Server Power Consumption Model

According to [97], saving 1 W of power at the CPU level could turn into 1.5 W of savings at the server level, and up to 3 W at the overall system level of the data center. The overall power consumption of the IT equipment can be reduced by reducing the power consumption of a single device or distributing the workload to the server clusters [98]–[100]. Thus, the power consumption model of the server is important to ensure the cost-effective operation of the data center. Regarding the applications of power consumption models, accuracy and simplicity are the main requirements, but they are contradictory and restricted [23]. As an example, a simple regression power consumption model of the serves is used to obtain the power consumption of the IT load, which is used further to assess the reliability and the voltage dips impacts in the IPCS [8], [101]. Meanwhile, the higher order regression models (i.e., quadratic and polynomial) of the server power consumption models are more complicated compared to the linear models. The complicated higher order regression model of server power consumption is used in [58] to improve the power efficiency of the servers by scheduling the task in a cloud interface, where the authors have focused on the accuracy of the model except the simplicity. Thus, the trade-off between accuracy and the simplicity of the consumption models of the servers depends on the application. The applications of the analyzed modeling approaches with advantages and disadvantages are summarized in Table 6.

B. INTERNAL POWER CONDITIONING SYSTEM MODEL

The IPCS of a data center consists of UPS, PDU, and PSU including the protection and power flow control devices (circuit breakers, automatic transfer switch, by-pass switch. etc.). The IPCS ensures the voltage quality and reliability of the power supply to the IT load section that guarantees the desired QoS [8], [18]. The IPCS of a data center consumes a significant amount of power during the voltage transformation process which is treated as power losses in [8], [52]. In a typical data center power hierarchy, a primary switchboard distributes power among multiple online UPSs. Each UPS supplies power to a collection of PDUs. A PDU feeds the IT load demand of the servers in a rack through PSUs located in the racks. A rack contains several chassis that host individual servers. The general representation of the IPCS is shown in Figure 6, which is explained in [8], [111], [112].

The PDU transforms the supplied high AC voltage to low AC voltage levels to distribute the power among the racks through the connected PSUs. The PDUs get the power from the UPS, while the UPSs are typically connected to the utility supply and backup generators, as shown in Figure 6. Depending on the region, the data center supplied voltage can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

TABLE 6: Applications, advantages, and disadvantages of the power consumption models.

Additive model	
Application	VM placement model considering energy cost, VMs placement cost, the communication cost Developing the entire data hall (server room) energy consumption model and simulate the proposed model in EnergyPlus, OpenStudio simulation interfaces [102]. Server power consumption and energy optimization [62], [63]
Advantages	Use widely to predict the power consumption of servers at large scale setup, which could be extended further to other section of data center level, as applied in [103].
Disadvantages	It requires various monitoring parameters to track the power consumption tend of different components of the servers (i.e., CPU, memory, fans, etc.)
Base Active model	
Application	To formulate a stochastic program that captures the data center-level load balancing, the server-level configuration, monitoring the Quality of Service (QoS) [104]. Managing the IT load of data center to make a balance between energy efficiency and QoS [105]. Power consumption estimation of CPU-dominated servers, medium utilization systems, cooling load calculation, and cloud computing management [73], [74].
Advantages	It only monitors one parameter that is the active power consumption of the server. The approach is suitable for relating the IT load power consumption with other load sections, specially to predict the cooling load section power consumption.
Disadvantages	Perform large prediction error for less CPU dominated systems, and limit to partial utilization regions and server types since it considers the undefined power consumption as the base power consumption of the IT load section.
Regression model	
Application	Energy consumption modeling of data center [106]. Optimize the energy consumption of mobile edge computing environment [107], [108].
Advantages	Easy to model the IT load section specially for applications like energy management and energy optimization for data centers.
Disadvantages	Most of the models are developed based on experiments with a specific type of experimental setup.
Utilization-based model	
Application	Optimized the VM allocation and resource prediction [109], dynamic consolidation of VM, enhance utilization of resources, and assessing the SLAs including an experimental results to validate the proposed model [110]. Overall consumption monitoring, energy optimization considering the power consumption of IT load section [58], [96], and other load sections [8], [52].
Advantages	It can capture the power consumption trend of different types of servers with different workloads, which is easy to related further with other application like energy optimization, power loss reduction, etc.
Disadvantages	Similar to the simple regression model thus power consumption models from this group are not suitable for predicting the power consumption of serves precisely.

vary from $480\,V_{AC}$ to $400\,V_{AC}$ that needs to be step down before distributing among racks [114]. The PDU works as a power converter to maintain the adequate voltage quality of the rack supply and the PSUs at racks rectify the supplied voltage for the servers using Switch Mode Power Supply Unit (SMPSU) [8]. The power electronic devices with high frequency switching like PDUs, incur a constant power loss as well as a power loss proportional to the square of the server load [52], as shown in (18). The PDU typically consumes 3% of its input power [52], [115]. As in current practice, all the PDUs remain connected with the supply system, which increases the idle loss of PDU [115]. The power loss coefficient of the PDU is represented by ϕ_{PDU} in (18) as explained in [28], [115].

The UPSs provide backup support during power supply interruptions up to some tens of minutes, voltage dips, and other disturbances originating upstream the UPS. Different types of UPS have been studied to evaluate the efficiency and performance for specific uses, while the *Online UPSs* are claimed to be the most-reliable choice for data center application because of the fast response time [114], [115]. Advancement has made recently to the internal topology of the online type UPS to improve the power quality [116], [117], efficiency [118], and performance [119], [120]. However, research on the power consumption or loss modeling of the UPS for data center application is very limited. The power

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

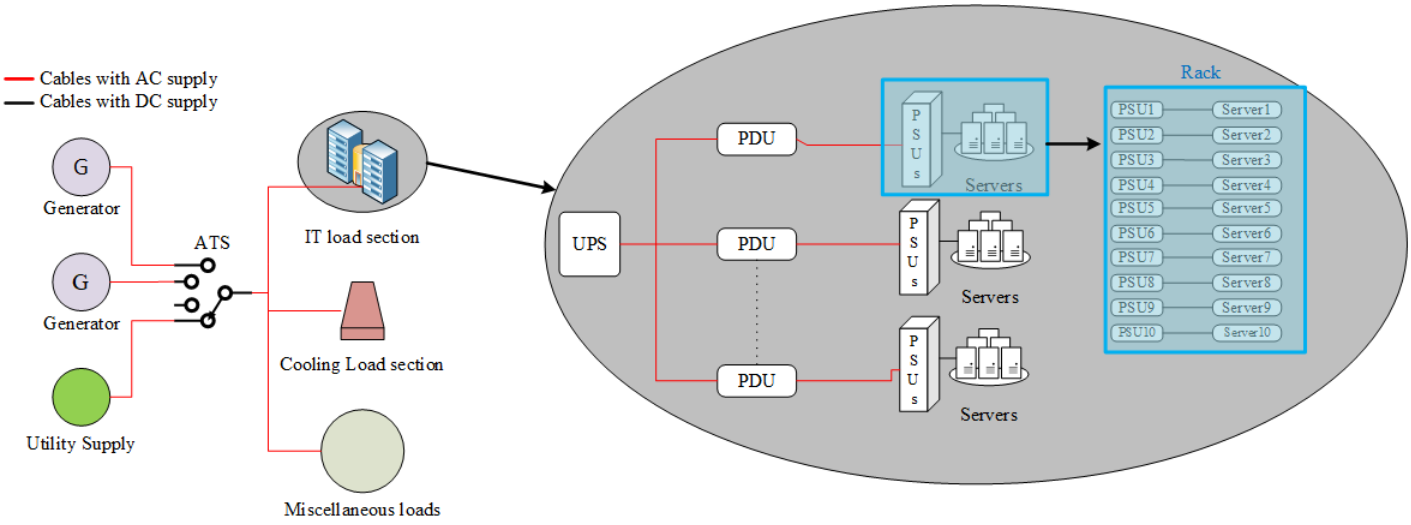


FIGURE 6: An example of the internal power conditioning system. [113]

consumption model of the UPS depending on the supplied IT load is proposed in [115] and later also used in [28], [52], [121]. The power consumed by the UPSs depends on the supplied power regardless of the topologies as shown in (19)

The power consumption of the PSU depends on its supplied power to the server [52], [53], [111]. The efficiencies of the PDU and the PSU are compared at different voltage levels of the data center in [114]. The efficiency of the PSU (87.56%), is less than the efficiency of the PDU (94.03%) for a 480 V_{AC} system in data center [114]. The efficiency is calculated based on the input and output power of each unit in [114]. However, the total power consumption of all PDUs is higher than the total power consumption of all PSUs [113] in the IPCS, because of the ideal power loss and the non-linear relation of the PDU's loading and power loss as shown in (18).

$$P_{PDU}^{Loss} = P_{PDU}^{idle} + \Phi_{PDU} \left(\sum_{servers} P_{server} \right)^2 \tag{18}$$

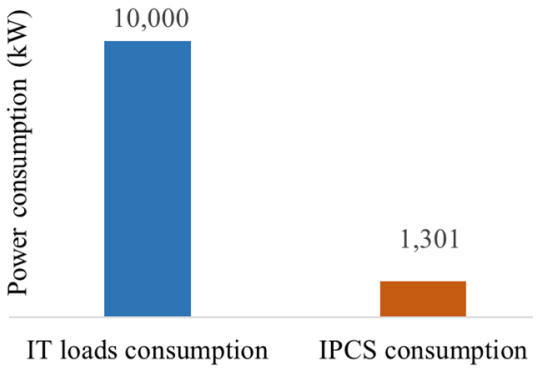
$$P_{UPS}^{Loss} = P_{UPS}^{idle} + \phi_{UPS} \sum_{PDU} P_{PDU} \tag{19}$$

A comparative study is shown in [52], where the performance of these devices in the IPCS has been evaluated in terms of consumed power by the IT loads in the data center. The PDUs are claimed to be the most power consuming equipment in the IPCS compared to UPS and PSU in [52], which can even lead to outages as explained in [8]. Due to the series power loss component in PDU that is represented by the square term in (18), the total power loss of the PDUs goes higher than the total power loss of the UPSs and PSUs [8], [52]. However, the efficiency of the PDU is compared with the UPS that shows the efficiency of PDU is higher than UPS [114], [115]. The power consumption of the devices in the IPCS in terms of percentage of the served IT loads for a hyperscale data center is shown in Figure 7. The analysis has been done based on the information that is presented in [112], [121] about the idle power consumption and the power loss coefficients of the UPS and PDU. The data center is considered with 10,000 servers with a rated power of 1 kW. A similar IPCS configuration is considered as shown in Figure 6, where each rack with 10 servers needs a PDU to distribute the power between the connected PSUs at the rack. Therefore, the data center is simulated with 10000 servers in 1000 racks that need 10 MW power for the servers. The racks are assumed to be supplied by 10 identical units of the UPS. The devices in the IPCS has consumed 1,301 kW of power to server 10 MW of the IT loads, which is around 13% of the power consumed by the IT loads, as depicted in Figure 7a. The power consumed by the devices in the IPCS is considered as the power loss in the IPCS [113]. In the assessed data center, the PDUs consume 7.3% of the power consumed by the IT loads while the UPS consumes 4.7% assuming the full computational loads for the servers, as shown in Figure 7b. The power loss of the PSU is assumed to be 1% of the supplied power to the servers since the power loss of the PSU is load dependent [113]. This analysis also shows the total power loss of the PDUs are more than the power loss of the UPSs as claimed in [8], [52].

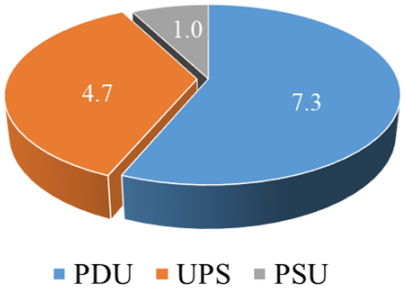
C. COOLING SECTION MODELS

The cooling and environmental control system is used to maintain the temperature and humidity of the data center. This sections mainly contains the Computer Room Air Cooling System (CRAC) unit, cooling tower, humidifiers, pumps, etc. to ensure the reliable coolant flow in the data hall. The highly-dense IT loads generate enormous amount of heat in data center, which is handled by the cooling load sections. The cooling loads ensure the environmental control and the IPCS ensures the power quality of the supply to the IT loads; while both of these load sections are needed to ensure uninterrupted service of the IT loads in the data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access



(a) Power consumed by the IT loads and the equipment in the IPCS.



(b) Power loss of the IPCS equipment as a percentage of IT load power consumption.

FIGURE 7: Power consumption of the IT loads and the equipment in the IPCS.

center. The cooling load section of the data center is the biggest consumer of power among the non-IT load sections followed by power conditioning system losses in a typical data center, as shown in Figure 4. The energy consumption models of the cooling section have various applications in data centers operation i.e., cooling section energy consumption management, optimization, generated heat utilization, thermal control, etc. The power consumption models of the cooling load section's components are essentially needed for the mentioned methods.

The power consumption of the cooling load section depends on multiple factors like the layout of the data center, the spatial allocation of the computing power, the airflow rate, and the efficiency of the CRAC [112], [122]–[124]. There are two major working components in the cooling section. (1) the CRAC unit, and (2) the chiller or a cooling tower.

1) CRAC Unit Models

The CRAC has recently drawn attention regarding efficiently handling the coolant flow in the data centers [125]–[127]. The heat generated from the servers, hence the IT loads in the data center are removed by the CRAC units installed in the server room. The cooling power that is consumed by the CRAC units is proposed in [125] as a function of supplied coolant temperature t_s and coefficient of performance C_{CoP}, as shown in (20). The authors use the HP CRAC model in [125] with a $C_{CoP} = 0.0068 \cdot t_s^2 + 0.0008 \cdot t_s + 0.458$, where t_s is the maximum temperature of the supplied coolant from the CRAC. The maximum efficiency of the CRAC unit can achieve by finding the maximum value of t_s that guarantees the reliable operation of the servers [125].

$$P_{CRAC_{cool}} = \frac{Q_{inlet}}{C_{CoP}} \tag{20}$$

Another power consumption mode of the CRAC system is presented in [128], where the CRAC is assumed to be equipped with variable frequency drivers (VFDs) which showed the following empirical relationship between individual CRAC unit power consumption P_{crac_i} and relative fan speed θ_i for the CRAC unit,

$$P_{crac_i} = P_{crac_i,100}(\theta_i)^{2.75}$$

The impact of racks arrangement, ambient temperature, outside temperature, and humidity on the power consumption of the CRAC unit is analysed in [129]. As explained in [129] the power consumption of CRAC is proportional to the volume of the airflow, f, and also depends on the heat generated by the servers, as shown in (21), (22). The required volume of air flow in a server room can be determined by $f = f_{max} \times \mathcal{U}$, where f_{max} is the maximum standard air flow ($14000\,\mathrm{m}^3/\mathrm{hr}$ for a 7.5 kW CRAC unit). The power required to transfer the heat P_{heat} from the server room is shown in (22), where the idle power of the CRAC unit, P_{CRAC}^{idle} can be considered as 7% to 10% of P_{sf}^{max}.

$$P_{heat} = 1.33 \times 10^{-5} \times \frac{P_{sf}^{max}}{\eta_{heat}} \times f \tag{21}$$

$$P_{CRAC} = P_{CRAC}^{idle} + P_{heat} \tag{22}$$

Recently a power consumption model of the CRAC is presented in [52] dominated by fan power, which grows with the cube of mass flow rate to some maximum ($P_{CRAC_{Dyn}}$), together with a constant power consumption for sensors and control systems ($P_{CRAC_{Idle}}$), shown in (23). Some CRAC units are cooled by air rather than chilled water or contain other features such as humidification systems, which are not considered here.

$$P_{CRAC} = P_{CRAC_{Idle}} + P_{CRAC_{Dyn}} f^3 \tag{23}$$

On the contrary, the power consumption of the CRAC in data centers is addressed in terms of thermal management in [126], [130], [131], where the authors relate the power consumption of the CRAC to the temperature of the data hall and the heat generated from the IT load section to optimize the power consumption of the cooling load section.

2) Chiller and Cooling Tower Power Consumption Model

There is not so much that has been done to address the chiller power consumption for the specific use case of data centers. The chiller plant removes heat from the warm coolant that returns from the server room. This heat is transferred to external cooling towers using a compressor. The chiller plant's

compressor accounts for the majority of the overall cooling power consumption in most data centers [128]. The power drawn by the chiller depends on the amount of extracted heat, the chilled water temperature, the water flow rate, the outside temperature, and the outside humidity. According to [18], the chiller's power consumption increases quadratically with the amount of heat to be removed and thus with the data center utilization. The size of the chiller plant depends on the maximum heat generated from the IT load section . According to the design practice the chiller should handle at least 70% of P_{sf}^{max} in order to provide sufficient cooling [132]. The chiller plant power consumption model is shown in (24). Another chiller power consumption model is given in [128], which depends on the power consumption of the refrigeration system P_r, as shown in (25). The constants α, β and γ are obtained by performing a curve fitting of several samples from the real data center.

$$P_{chiller} = 0.7 \times P_{sf}^{max} \left(\alpha \mathcal{U}^2 + \beta \mathcal{U} + \gamma\right) \tag{24}$$

$$P_{chiller} = P_r / \eta \tag{25}$$

where η and $\mathcal{U}$ are the efficiency of the chiller system and the average utilization of the servers in the IT load section.

3) Power Consumption Model of the Cooling Section

The additive power models are common for modeling the data center's cooling section power consumption like IT load section. An additive model for power consumption of the cooling system of the data center is presented in [133] and shown in (26). The power consumption model includes the CRAC fan, refrigeration by chiller units, pumps of the cooling distribution unit, lights, humidity control, and other miscellaneous items [133]. P_{rf} corresponds to the total power consumption of the cooling system for a raised floor architecture, known as a refrigeration system. P_{CRAC} is the power consumed by computer room air conditioning units. P_{cdu} denotes the power dissipation for the pumps in the cooling distribution unit (CDU) which provides direct cooled water for rear-door and side-door heat exchangers mounted on the racks. P_{misc} is the power consumed by the miscellaneous loads in the cooling system. This model is almost equal to the model of raised floor cooling system power consumption described in [128].

$$P_{rf} = P_{CRAC} + P_{cdu} + P_{misc} \tag{26}$$

The total power consumption of the CRACs and total power consumption of the CDUs could be expressed as follows, where i and j corresponds to the number of CRAC and CDU units, respectively. P_{CRAC} and P_{cdu} are the total power consumption of the CRAC and CDU units.

$$P_{CRAC} = \sum_i P_{CRAC_i}, \text{ and } P_{cdu} = \sum_j P_{cdu_j}$$

A summery of the data center load modeling analysis with the references is given in Table 7.

IV. REVIEW OF THE RELIABILITY MODELING OF DATA CENTERS

Data centers should be environmentally controlled and equipped with power conditioning devices to ensure the reliable operation of the IT loads including servers and network devices. Data center operators take every possible measure to prevent deliberate or accidental damage to the equipment in the data center so that the load sections could ensure a high degree of reliability in operation. By definition, reliability is the probability of a device or system performing its function adequately under specific operating conditions for an intended period of time [134], [135]. Here the degree of trust is placed in success based on past experience, which is quantified as the probability of success for a mission oriented system like a data center in this case. This reliability definition considers only the operational state of the component or system without any interruptions. Meanwhile, the probability of finding the component or system in the operating state is known as "availability", which is used as a reliability index for a repairable system [135]. In this case, the components in the data center load sections are repairable that also includes the replacement process, therefore the availability index is widely used in data center reliability modeling [136].

The data center industry has come to rely on "tier classifications" introduced by the Uptime Institute as a gradient scale based on data center configurations and requirements, from the least (Tier 1) to the most reliable (Tier 4) [136]. The Uptime Institute defines these four tiers of data centers that characterize the risk of service impact (i.e., unavailability and downtime) due to both service management activities and unplanned failures [137].

A. TIER CLASSIFICATION OF DATA CENTERS

The core objective of the tier classification of data centers is to make a guideline of the design topology that will deliver desired levels of availability as dictated by the owner's business case, which is introduced by the Uptime Institute [136], [137]. The tier of the data center is determined by the availability of the IPCS including the utility and backup generator supply [136], [137]. The Uptime Institute is the pioneer in researches to standardize the data center design and describe the redundancy of its underlying power supply systems. According to The Uptime Institute's classification system, the internal infrastructure of data centers has evolved through at least four distinct stages in the last 40 years, which is used for the reliability modeling and known as "Tiers of Data center" [136]–[138]. As of April 2013, the Uptime Institute had awarded 236 certifications for building data centers around the world based on the tier classification [139]. This is a combination of quantitative and qualitative classification approach, as depicted in Figure 8. The combination of these two approaches is used by the Uptime Institute for tier certification, however, the reliability assessment approach depends on the data center owner's business cases, which is discussed further in Section IV-D. The tier classification system evaluates data centers by their capability to allow

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

TABLE 7: The summary of the reviewed topics in data center load section modeling with the references

Reviewed topics	References
Application of the power consumption models	10, 25, 28, 33, 51-61
IT load models	23, 24, 39-41
Server consumption model	8, 23, 53, 58, 60, 62-68, 70-96
1. Additive power model	62-68, 70, 72
2. Baseline - Active (BA) power model	53, 73-76
3. Regression model	23, 74, 75
a. Simple regression model	60, 73, 78-83
b. Multiple regression model	73, 84-90
c. Non-linear model	58, 78, 83, 91, 92
4. Utilization based model	58, 78, 81, 83, 93-96
Importance of server power consumption model	8, 23, 97-101
Applications, advantages, and disadvantages of servers power consumption models	52, 62, 63, 73-74, 102, 110
Internal Power Conditioning System Model	8, 18, 52, 53, 111-115, 121
Cooling Section Models	112, 122-124, 128, 133
1. CRAC unit model	52, 125-131
2. Chiller and cooling tower consumption model	18, 128, 132

maintenance and to withstand a failure in the power supply system. Tier I (the least reliable) to Tier IV (the most reliable) are defined depending on the redundant components in the parallel power supply path to the critical load sections. However, the deterministic approach used in [136], [139] to calculate the availability for different tiers has ignored the outage probability of the grid supply, different failure rates of the IPCS components, and random failure modes in the power supply paths. The specification and redundant options from [136]–[138] are summarized in Table 8.

The availability of the data center for different tiers that are given in Table 8 are criticized in [140]. The availability of the data centers that are shown in [140] are less than the former ones. Due to considering more detailed failure possibilities in the data center internal infrastructure the risk of failure increases, hence the availability decreases for the studied system in [140]. Therefore, the redundancy in the power supply path can not only improve the availability of the data center; the availability could degrade due to common mode failures, which demands statistical data for further research. Although a crude framework and design philosophy that is provided in [136]–[138] is still useful, the results are presented based on some assumptions, as follows:

- The fault tolerance of the tiers does not solely depend on the redundancy of the power supply path because there is a possibility to have common mode failures. The impact of the common mode failures in rack-level PSUs on the availability of the servers are presented in [113].
- The studies only consider the single point of failures in specific critical output distribution points like PDUs and provide a solution to use dual corded PDUs in Tier IV data center. However, it is argued in [8] that the dual corded PDUs also could fail to supply the required power to the servers because of power supply capacity shortage.
- The articles have followed a deterministic approach with constant failure and repair rates of the components to assess the availability of small IPCSs, while the IPCSs in real data centers are large and complex with a high number of uncertainties to have component outages at different levels in the IPCS.

B. FACTORS TO CONSIDER FOR THE RELIABILITY MODELING IN DATA CENTERS

The most important factor for assessing the reliability of the data center is the *failure* of the components in the system. Arno et al. has formulated an example in [136] as follows:

"*If the UPS in the power supply system fails and all the connected loads for the data center lose power, that would obviously be a "failure." But what about one* 20 *A circuit breaker trips and one rack of equipment losing power? Is that a "failure" for the data center?*" [136]

According to the definition of failure given in Chapter 8 of the IEEE Gold Book, Standard 493-2007 [141], "the failure is the loss of power to a power distribution unit (or UPS distribution panel in case of the data center)." Thus the loss of an entire UPS would impact the overall mission of the connected facility that is a failure of the data center by definition. However, if a circuit breaker trips and the connected racks lose power then it will not be considered as a failure of the data center, rather the servers at the racks are considered as failed or unavailable for operation. Therefore, the first step of any reliability analysis is to define the "failure state" of the studied system. A similar explanation is presented in [7] about "error and failure" for a cloud system, where the term "failure" is used for fatal faults in the system that are irreparable and catastrophically impact the system operation. However, "errors" degrades the system performance (i.e.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

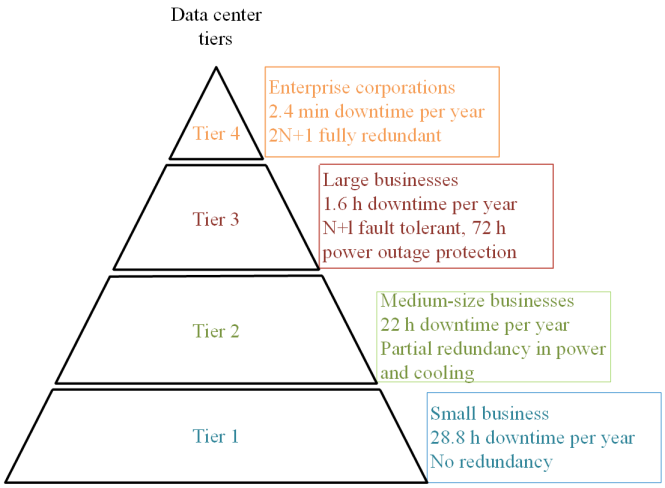


FIGURE 8: Data center tier classification.

TABLE 8: Overview of tier classification requirement [136]–[138]

	Tier I	**Tier II**	**Tier III**	**Tier IV**
Utility Supply (Connection point)	Single point	Single point	Single point	Dual
No of backup generator	Optional	N	N+1	2N
Backup system (UPS)	N	N+1	N+1	2N
Maintenance	outage for maintenance	outage for maintenance	concurrently maintainable	fault tolerant
Availability	0.999947	0.9999512	0.9999791	0.9999976
Availability (Nr. of nine's)	4	4	4	5

latency, decreasing throw put) since the errors can be solved automatically and the system can recover to the initial state [7]. Additionally, the mentioned failure definition in the IEEE Gold Book, Standard 493-2007 [141] contradicts with the tier classification of data center, which shows failure with degraded performance mode is needed to be defined for data centers. The reliability analysis in [113] is an example in this regard. The failures of the rack-level PSUs are considered in [113] to assess the adequacy of the computational resources, hence the degraded performance of the data center.

C. RELIABILITY INDICES AND METRICS USED FOR RELIABILITY MODELING IN DATA CENTERS

In this section the reliability indices and metrics that are used in literature for data center reliability modeling are analyzed in three groups depending on the load sections. The applications of the reliability indices in reliability modeling differs for different load sections because the interpretation of the reliability assessment outcomes are not similar for the load sections, as mentioned in Section IV-B.

1) Reliability Indices for IT Loads and Services

The indices that are used for IT load section could be classified into two groups, (1) the indices that are related to the IT performance and services, and (2) the indices related to the readiness of the IT load section in data center. The QoS is a key indicator to assess the performance of the data center, which also includes Key Quality Indicators (KQI) and Key Performance Indicators (KPI) for the IT services provided by the data center as explained in [7]. These indices are used for IT service monitoring and computational capacity management in data center. The "service reliability" and "service availability" indices are used to maintain SLA with the client or user of the data center. In other words, the service availability or reliability characterizes the readiness of a data center system to deliver the promised IT service to a user. The readiness of a system is commonly referred to as being "up" [142]. Mathematically, the service availability is estimated as given in (27). The "service availability" index is used in [8] for addressing the reliability of the IT loads or the servers at rack-level. The authors have also shown the server "outage probability" as a reliability index that varies with the increasing power losses in the IPCS in the data center.

$$A_{Service} = \frac{t_{up}}{t_{up} + t_{down}} \tag{27}$$

where, $A_{Service}$ is the service availability. t_{up} and t_{down} are the uptime and downtime of the system, respectively. Apart from the probability of outages, the "service reliability" is also emphasized in [142] since the probability to fulfill the service requests without latency is characterized by this index. Importantly session-oriented services in data centers measure both the probability of successfully initiating a session with service, called "accessibility", and the probability that a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3125092, IEEE Access

session delivers service with promised QoS until the session terminates, called "retainability" [142]. In this regards the Defects Per Million Operation (DPM) is an index that measures the failed operation per million operations to assess the system reliability, as given in (28).

$$R_{DPM} = \frac{O_f}{O_a} \times 1,000,000$$
$$r_{Service} = 100\% - \frac{R_{DPM}}{1,000,000} \tag{28}$$

where, R_{DPM}, O_f, and O_a are the defects per million operation, the number of failed operation, and attempted operation, respectively. The service reliability is represented by $r_{Service}$.

Another index named "Service latency" is mentioned with importance to assess the system reliability specially for edge and internet data centers in [143], [144]. Transaction latency directly impacts the quality of experience of end users; according to [142], 500 millisecond increases in service latency causes a 20% traffic reduction for Google.com, and a 100 millisecond increase in the service latency causes a 1% reduction in sales for Amazon.com. The service availability index is explained considering the average CPU load level, hence computational workloads, and CPU hazard function, with a new index "load-dependent machine availability" in [145]. Similar load-dependent reliability indices named "average performance" and "average delivered availability" are proposed in [146].

The basics of the QoS and the service reliability indices are similar; mostly based on the indicators of the service availability and the IT system performance. The IT system performance indicators are modeled in different ways, such an example is given in (28).

Apart from the mentioned QoS oriented reliability indices, there are other indices i.e., Mean Time Between Failure (MTBF), Mean Time To Repair (MTTR), availability, reliability that are used in reliability modeling for the physical components of the IT load section [147]–[149]. A similar reliability index called "loss of workload probability (LOWP)" is proposed based on the server outages probability at the rack-level in [113]. The risk of server outages due to electrical faults and the consequent voltage dips are analyzed in [101].

Additionally, the IT load performance based SLA-aware indices are also used for the software-based solutions in data centers. The SLA-aware indices i.e., Performance Degradation Due to Migration (PDM), Service Level Aggrement Violation (SLAV) are applied to evaluate the performance of the IT loads with consolidated workloads in the cloud system [150], [151].

2) Reliability Indices for IPCS Section

The indices that are used to assess the reliability of the IPCS in the data center are compiled with a logical explanation in [136]. The authors specify five different reliability indices in [136] i.e., MTBF, MTTR, availability, severity and risk (measured in terms of financial losses caused by the failure) for assessing the reliability of the IPCS in data centers. These indices are significantly impacted by the definition of "failure" that is used for the studied system architectures as explained in Section IV-B. The indices are also impacted by the size of the facility and the number of the critical loads used in the studied models [114]. A different reliability assessment approach is explained in [8], where authors showed the power supply capacity shortage probability of the PDUs due to increasing power losses in the IPCS that eventually results in server outages, hence failure in the IT loads. The index named "outage probability" is used for PDUs to relate with the service availability of the IT loads [8].

There are also reliability studies focused on the IPCS components (i.e., UPS, PDU, and PSU). These articles are out of scope of this review since the component based research is focused on lifetime enhancement, cost-effectiveness, and energy-efficiency of the particular component but not focused on data center applications.

3) Reliability Indices for Cooling Section

A reliability evaluation method for a hybrid cooling system combining with a lake water sink for data center is presented in [152], where the operational availability index $A_O(\infty)$ is used for repairable system components. In [152] the operational availability index is defined as the probability that the system will be in the intended operational state, and mathematically expressed as a function of system's failure rate λ_{sys} and repair rate μ_{sys}, as given in (29). Another reliability index called functional availability $A_f(\infty)$ is also used based on predicted server room temperature and servers' working conditions in [152], as given in (30). As explained in [152], the overall functional availability of the data centers cooling system is mainly determined by the operational availability, heat density, heat transfer characteristics of room temperature, start-up time of cooling system and repair time of cooling system failure.

Similar functional condition based analysis has been done for the data center air-conditioning system based on the air-conditioning power supply capacity [153].

$$A_O(\infty) = \frac{1}{\frac{\lambda_{sys}}{\mu_{sys}} + 1} \tag{29}$$

$$A_f(\infty) = A_o(\infty) \times (1 - p_{us}) + (1 - A_o(\infty)) \times p_t \tag{30}$$

where $A_O(\infty)$ and $A_f(\infty)$ are the operational and functional availability of the cooling system. The failure rate and repair rate of the cooling system are λ_{sys} and μ_{sys}, respectively. p_{us} is the probability of room temperature out of intended range when the system is under operation state. p_t is the probability of intended value of room temperature when the cooling system fails.

A different reliability modeling approach has been applied in [9], where authors emphasized on dependability of the cooling system since dependability is related to both fault tolerance and reliability. The reliability importance (I_i) and

reliability-cost importance (C_i) indices are used in [9], as given in (31)

$$I_i = R_s(U_i, \mathbf{p}^i) - R_s(D_i, \mathbf{p}^i)$$
$$M_i = I_i \times (1 - \frac{C_i}{C_{sys}}) \quad (31)$$

where, I_i is the reliability importance of component i; $\mathbf{p}^i$ represents the component reliability vector with the i^{th} component removed; D_i and U_i represent the failure and up state of i component, respectively. C_i is the acquisition cost of the component i and C_{sys} is the system acquisition cost.

Apart from the mentioned indices, the typical indices like availability based on MTBF, MTTR, failure, and repair rates are widely used for reliability modeling of the cooling load section of data center [154], [155]. It is important to mention that the research on data center cooling system reliability is not adequately addressed yet, while the cooling infrastructure for commercial buildings has already drawn the interest of the researchers intensively in the last decade. The research on reliable cooling infrastructure of the data center is much needed since the temperature sensitivity of the data center's server hall needs to be compared to the other building facilities [154]. One of the very few articles that have critically evaluated the reliability of the cooling system of data center recently is [154].

D. METHODOLOGIES USED FOR RELIABILITY MODELING

Different research methods have been used for reliability modeling of the data center's load sections individually and also the data center as a complete system. All the proposed methodologies could be classified in two groups i.e., analytical research group, and simulation based research group [156].

1) Analytical Approaches for Reliability Assessment

The applications of analytical approaches like Reliability Block Diagrams (RBD) and fault tree analysis are very common in data center reliability modeling because of the simplicity and less requirement of the computational capacity. One of the earliest of such research was published in 1988 [157], where the authors analyzed and compared the unavailability of the distributed power supply system of a telecommunication control room with the centralized power distribution system. A similar analytical approach has been explained in [158], where the reliability of the typical Alternate Current (AC) distribution system is compared with the Direct Current (DC) power distribution system in data centers using RBD. The failure of the power distribution system is only considered without considering the failures of the IT loads in [158], [159], while the availability of the IPCS considering the failure probability of the IT loads including PSUs are presented in [8]. Depending on the voltage level in the IPCS the reliability of different IPCS structures are evaluated using RBD in [160], [161]. The reliability modeling of the computation resource infrastructures (IT load section) of data centers has been conducted using RBD model in [162], [163]. A similar type of analysis is presented in [164], where the authors have used the directed and undirected graphs using minimum cut sets. The analytical approach is also applied to evaluate the reliability of the data center's network topologies by applying the concept of cut set theory [165], [166], and optimizing the resource allocations for reliable networks [167]. The analytical approach i.e., the RBD, stochastic Petri net and energy flow model are used for reliability assessment of the IPCS in [168]. An extended RBD model is proposed in [169] that can consider the dependency of the IPCS components' reliability on the overall reliability of the IPCS. The proposed model is called Dynamic RBD, which is further compared with colored Petri net model in order to perform behavior properties analysis that certifies the correctness of the proposed model for IPCS reliability, as explained in [169]. The fault tree analysis technique is used to estimate the failure rates, MTBF, MTTR, and reliability of different UPS topologies in [170]–[172].

The RBD is also used for data center cooling system reliability analysis in [9], [152]. The availability of a water-cooled system is evaluated using maximum allowable downtime in the proposed RBD model in [152], while the RBD and stochastic Petri net model are used for quantification of sustainability impacts, costs, and dependability of data center cooling infrastructure in [9]. A comparison of data center sub-systems' reliability is presented in [148], where the reliability of the network, electric, and thermal system of the data center is modeled using the failure modes effects with criticality analysis (FMECA) and energy flow model (EFM). The proposed methodologies in these mentioned articles are evaluated using the components' statistical failure and repair data. There are common sources of these data for industrial applications like [173], [174]. However, the infrastructures of the data centers are more critical than typical buildings and industries as argued in [154]. The statistical data of the data center's component failure is needed for further research to improve the competent reliability in data center application. There is a publicly available data set that publishes the failure and repair times of the servers [175], while the failure and repair data of other components (i.e., PDUs, PSUs, cooling devices) are not part of any publicly-available set of data. The data center operator's tendency to hold the confidentiality and secrecy of the internal information of the data centers are the main reasons behind the lack of such data sets [176].

2) Simulation-based Approaches of Reliability Assessment

Along with the analytical models, the probabilistic modeling approaches are also common for data center reliability assessment. The state space models including Markov model and Markov chain Monte Carlo (MCMC) are used for reliability modeling of large scale and repair-able systems, therefore the application of Markov models have become popular for reliability modeling of data centers recently [177], [178]. To avoid the time-variant non-linear state space model in

Markov model, the failure and repair rate of the components of the studied systems are assumed to be constant. The failure and repair rate could be constant for a component if the aging effect is ignored considering a constant failure rate [135]. Therefore, the simulation based reliability models for assessing the reliability of the data centers are widely used in research nowadays.

Monte Carlo is one of the most used simulation-based approaches for data center reliability modeling. The Monte Carlo simulation approach is mostly used to generate time-dependent failure and repair events of the system components using probability distribution function, and observe the overall system performance based on the stochastic data [156], [179], [180]. The Monte Carlo simulation method is also used for reliability modeling of the components that are used in data centers i.e., UPS [181], [182], optical network system [180]. In simulation-based approaches the failure model of the system's component is important since the simulated result of overall reliability can vary depending on the failure mode, especially for the high reliability application like the data center [8]. As an example, the availability of the Tier IV data center is required to have five to six 9's, which means very few failure events will be observed in a million stochastic events. Therefore, accuracy in failure mode consideration and component's failure modeling are important in the simulation-based approaches for reliability modeling of data centers. Apart from the number of samples and failure mode of the components, the probability distribution functions of the failure and repair events of the components in the studied system also play a crucial role in the simulation-based approaches in case of reliability modeling. The probability distribution functions and the applications of the distribution functions for reliability modeling of the servers in the data center are analyzed in [183]–[185]. The distribution function of the failure and repair time of the network devices and other server components i.e., hard-disk, memory, and network cards are presented in [176], which is further used for reliability modeling of the overall system. Besides Monte Carlo, stochastic petri nets [32], and Markov chain Monte Carlo (MCMC) [186] are also used for reliability modeling of the data center.

E. DEPENDABILITY OF THE DATA CENTER LOAD SECTIONS AND SUB-SYSTEMS

The dependability of a system is defined as the ability of the system to deliver the service that can justifiably be trusted [187]. Alternatively, providing the criterion for deciding if the service is dependable is the dependability of a system [188]. As an example, the dependence of system A on system B represents the extent to which system A's dependability is (or would be) affected by that of system B. The dependability of a system can be represented by attributes i.e., availability, reliability, integrity, maintainability, etc [188]. This section analyses the dependability of the data center sub-systems and load-section since the service availability of the data center depends on the continuity of the services provided by the components of the sub-systems, as explained in Section IV-C.

1) Dependability on the Cooling Load Section

A dependability analysis of data center sub-systems has been presented in [148], where the authors considered the availability of the three major sub-system (electric, cooling, and network) and also evaluate the impact of sub-system's availability on the data center reliability. The impacts of the electrical and thermal subsystem's availability on the overall reliability of the data center are presented in [32]. The impact of the ambient temperature on the overall reliability and energy efficiency of data centers has been analyzed in [147]. It has shown that the battery life in the IPCS is reduced by 50% due to increase in operational temperature by 10^o C; while the passive elements in the servers like capacitors' reduce the life time by 50% for 10^o C increment in the temperature [147]. The author in [147] has also concluded that increasing the data hall temperature improves the energy efficiency but it impacts the reliability of the servers and the PSUs in the IPCS. The power consumption of the cooling loads depends on the servers arrangement in data hall, hence, dense server arrangement causes high energy consumption by the cooling loads [32]. Additionally, the network and storage latency increases due to have overloaded cooling loads and have more un-utilized or idle servers, which also impact the overall relaiblity of the data center placement strategies, as explained in [32]. However, these articles have not considered the power losses of the IPCS to evaluate the reliability of the data center.

2) Dependability on the IPCS

The impacts of the power losses on the service availability of the IT loads of the data center are analyzed in [8]. The service availability of the severs, hence the IT services of the data center is quantified considering the total power losses in the IPCS. According to [8], the server outage possibility could be 20% of the installed capacity from the system because of the power loss of the PDUs in the IPCS. Moreover, the impacts of electrical faults and unwanted outages in the IPCS on servers' outages in data centers are presented in [101], [113]. The faults in the IPCS causes voltage dips and leads to trip the PSUs and the servers, as explained in [101]. The amount of workload that can not be handled for such unwanted failures are quantified in [113], since the failure could cause almost 33% of the insulated servers to be out of order in extreme cases [101]. The reliability-centric dependability analysis is further extended to control the computational resources to reduce the overall power consumption, hence the number of servers in the data center by balancing and scheduling the workloads in [189], [190]. The term "right-sizing" is used in this regards, although right-sizing is also used for reducing the number of idle servers based on data traffic and negotiated SLA in [191]. The number of active servers is optimized by workload consolidation through virtualization as proposed in [192], [193]. A different approach is presented in [113], where the authors address the required number of servers

per rack considering the workloads and stochastic failure of PSUs in the IPCS. The broader aim of these analyzed articles is to improve the reliability and energy efficiency of the data center, here the consumption models of the load sections are necessarily used. The energy consumption models are used either for internal structural modification to reduce the power losses [189], [190] or for allocating servers to minimize the consumption [191]. Therefore, the trade-off between energy efficiency and reliability enhancement in the data center is important to be considered in data center operation, where the energy consumption models of the data center load sections are often necessary for data center reliability assessments.

A summery of the data center reliability analysis with the references is given in Table 9.

V. LIMITATIONS AND FUTURE WORKS

This paper does not consider the energy management techniques that are used for improving the efficiency of the data center, whether the authors are focused on the energy consumption models of the data center's major components. Moreover, the adaptation of the sustainable and green energy sources in data centers are the novel challenges in data center operation. The impacts of the green technologies i.e., renewable energy generation and free cooling techniques in the energy consumption modeling approaches are not addressed in this paper. The adaptation of the sustainable energy sources in the data centers and its impacts on the reliability of the data center will be analyzed in future.

The detailed mathematical models of different simulation methods i.e., Monter Carlo, Markov Chain Monte Carlo, stochastic petri nets, etc. are not included in this review. These models are used as a tool for data center reliability assessment. This paper only reviews the applications of these models in the simulation-based reliability assessment techniques of the data center without considering the mathematical models, which could be considered for further study.

VI. CONCLUSION AND RECOMMENDATIONS

Being the backbone of today's information and communication technology (ICT) developments the energy-efficiency and higher reliability of the data centers are needed to be ensured in data center operation. In this paper the energy consumption modeling and reliability modeling aspects of the data centers are reviewed. The review has revealed the state-of-the-art of the aforementioned topics and the research gaps that exist in published review articles. This paper contributes to fill the research gaps related to data center energy consumption modeling by analyzing the energy consumption models of data center load sections, which will ease the models application in further research. It is worth mentioning that this paper reviewed the data center's reliability assessment models and methodologies for the first time, which also shows the existing research gaps as recommendations. The identified research gaps, hence the recommendations based on the analysis of data center reliability assessment review are needed to be filled by the future researcher to ensure the adaptation of new equipment and technologies in the data center. Additionally, it has been revealed that the energy consumption models of the data center components are often necessary for the data center reliability models, although the energy consumption models have also other applications (summarized in Table 1) for the data center energy management.

This paper recommends based on the review of the energy consumption models of data center components to emphasize more on the availability of the energy consumption model parameters and variables than the accuracy for applying in the research. The higher accuracy of such models often makes the application complicated and could not contribute much to the improvement of the proposed methodology. Additionally, the lack of research on the energy consumption modeling of the internal power conditioning system's (IPCS) equipment is identified in this review. The total power consumption of the IPCS could be rich up to 10% of the total demand of the data center, which could also cause outages and reliability issues in data centers. This review also contributes to show the relation between the power consumption and the reliability of the data center, and concludes more research should be conducted to reduce the power consumption specially in IPCS section, as a recommendation.

The data center reliability modeling aspects are reviewed in this paper that shows a need of standard code for data center operation along with existing tier classification, which is mentioned as recommendation. The analysis also contributes to show the state-of-the-art of the analytical and simulation-based reliability modeling approaches that could help future researchers to choose suitable models based on application. The analysis has shown the need of statistical failure and repair data of the data center components that is rarely available due to the operator's lack of willingness to share. Therefore, it is recommended to publish the component's statistical failure and repair data so that it could be used for further research. It is also a recommendation of this paper to give more focus on improving the cooling section reliability analysis and analyze the dependency of the data center's overall reliability on other load sections more in details.In its essence, this review has identified a few research gaps and a number of recommendations for the researcher to continue the research and improve the understanding of the data center's energy consumption and reliability modeling.

List of Abbreviations

AC Alternate Current.

BA Base Active.

CDN Content Distribution Network.
CPU Central Processing Unit.
CRAC Computer Room Air Cooling System.

DC Direct Current.
DPM Defects Per Million Operation.

TABLE 9: The summary of the reviewed topics in data center reliability analysis with the references

Reviewed topics	References
Tier classification of data centers	8, 113, 136-140
Factors to consider for the reliability modeling in data centers	7, 136, 141
Reliability indices and metrics used for reliability modeling in data centers	7-9, 101, 113, 136, 142-154
1. Reliability indices for IT loads and services	7, 8, 101, 113, 142-151
2. Reliability indices for IPCS section	8, 114, 136
3. Reliability indices for cooling section	9, 152-154
Methodologies used for reliability modeling	9, 135, 148, 152-186
1. Analytical approaches for reliability assessment	9, 148, 152, 154, 157-176
2. Simulation-based approaches of reliability assessment	135, 156, 176, 178-186
Dependability of the data center load sections and sub-systems	8, 32, 101, 113, 147-148, 187-193

I/O Input and Output devices.
ICT Information and Communication Technology.
IDC International Data Corporation.
IPCS Internal Power Conditioning System.
IT Information Technology.

KPI Key Performance Indicators.
KQI Key Quality Indicators.

LLC Last Level Cache.

MTBF Mean Time Between Failure.
MTTR Mean Time To Repair.

PDM Performance Degradation Due to Migration.
PDU Power Distribution Unit.
PSU Power Supply Unit.
PUE Power Usage Efficiency.

QoS Quality of Service.

RBD Reliability Block Diagrams.

SLA Service Level Agreements.
SLAV Service Level Aggrement Violation.
SLR Systematic Literature Review.
SMPS Switch Mode Power Supply Unit.

UPS Uninturrupted Power Supply.

VM Virtual Machine.

REFERENCES

[1] Munan Li and Alan L. Porter. Can nanogenerators contribute to the global greening data centres? Nano Energy, 60:235–246, 2019.
[2] Charles J. Corbett. How Sustainable Is Big Data? Production and Operations Management, 27(9):1685–1695, sep 2018.
[3] Rich Miller. The sustainability imperative: Green data centers and our cloudy future. Technical report, Data center Frontier.
[4] Yanan Liu, Xiaoxia Wei, Jinyu Xiao, Zhijie Liu, Yang Xu, and Yun Tian. Energy consumption and emission mitigation prediction based on data center traffic and PUE for global data centers. Global Energy Interconnection, 3(3):272–282, jun 2020.
[5] Jakob Dybdal Christensen, Jens Therkelsen, Ivo Georgiev, and Henrik Sand. Data centre opportunities in the Nordics An analysis of the competitive advantages. 2018.
[6] Jeffrey Rambo and Yogendra Joshi. Modeling of data center airflow and heat transfer: State of the art and future trends. Distributed and Parallel Databases, 21(2):193–225, 2007.
[7] Eric Bauer and Randee Adams. Reliability and availability of cloud computing. John Wiley & Sons, 2012.
[8] Kazi Main Uddin Ahmed, Manuel Alvarez, and Math H.J. Bollen. Reliability analysis of internal power supply architecture of data centers in terms of power losses. Electric Power Systems Research, 193:107025, 2021.
[9] Gustavo Callou, Paulo Maciel, Dietmar Tutsch, and Julian Araujo. Models for dependability and sustainability analysis of data center cooling architectures. In Proceedings of the International Conference on Dependable Systems and Networks, 2012.
[10] Yulia Berezovskaya, Chen Wei Yang, Arash Mousavi, Valeriy Vyatkin, and Tor Bjorn Minde. Modular model of a data centre as a tool for improving its energy efficiency. IEEE Access, 8:46559–46573, 2020.
[11] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. Systematic literature reviews in software engineering - A systematic literature review. Information and Software Technology, 51(1):7–15, jan 2009.
[12] Hongjie Lu, Zhongbin Zhang, and Liu Yang. A review on airflow distribution and management in data center. Energy and Buildings, 179:264–277, 2018.
[13] Wen-Xiao Chu and Chi-Chuan Wang. A review on airflow management in data centers. Applied Energy, 240:84–119, 2019.
[14] Chaoqiang Jin, Xuelian Bai, and Chao Yang. Effects of airflow on the thermal environment and energy efficiency in raised-floor data centers: A review. Science of The Total Environment, 695:133801, 2019.
[15] Hafiz M Daraghmeh and Chi-Chuan Wang. A review of current status of free cooling in datacenters. Applied Thermal Engineering, 114:1224–1239, 2017.
[16] Jiacheng Ni and Xuelian Bai. A review of air conditioning energy performance in data centers. Renewable and sustainable energy reviews, 67:625–640, 2017.
[17] Weiwen Zhang, Yonggang Wen, Yew Wah Wong, Kok Chuan Toh, and Chiu-Hao Chen. Towards joint optimization over ict and cooling systems in data centre: A survey. IEEE Communications Surveys & Tutorials, 18(3):1596–1616, 2016.
[18] Eduard Oró, Victor Depoorter, Albert Garcia, and Jaume Salom. Energy efficiency and renewable energy integration in data centres. Strategies and modelling review, feb 2015.
[19] Eduard Oró, Victor Depoorter, Noah Pflugradt, and Jaume Salom. Overview of direct air free cooling and thermal energy storage potential energy savings in data centres. Applied thermal engineering, 85:100–110, 2015.
[20] Alfonso Capozzoli, Marta Chinnici, Marco Perino, and Gianluca Serale. Review on performance metrics for energy efficiency in data center: The role of thermal management. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8945:135–151, 2015.
[21] Jianxiong Wan, Xiang Gui, Shoji Kasahara, Yuanyu Zhang, and Ran Zhang. Air Flow Measurement and Management for Improving Cooling and Energy Efficiency in Raised-Floor Data Centers: A Survey. IEEE Access, 6:48867–48901, aug 2018.

[22] Kofi Owura Amoabeng and Jong Min Choi. Review on cooling system energy consumption in internet data centers, dec 2016.

[23] Chaoqiang Jin, Xuelian Bai, Chao Yang, Wangxin Mao, and Xin Xu. A review of power consumption models of servers in data centers. Applied Energy, 265(March):114806, 2020.

[24] A review on energy efficiency and demand response with focus on small and medium data centers. Energy Efficiency, 12(5):1399–1428, 2019.

[25] Sami Alkharabsheh, John Fernandes, Betsegaw Gebrehiwot, Dereje Agonafer, Kanad Ghose, Alfonso Ortega, Yogendra Joshi, and Bahgat Sammakia. A brief overview of recent developments in thermal management in data centers. Journal of Electronic Packaging, Transactions of the ASME, 137(4), 2015.

[26] Chang Ge, Zhili Sun, and Ning Wang. A survey of power-saving techniques on data centers and content delivery networks. IEEE Communications Surveys & Tutorials, 15(3):1334–1354, 2012.

[27] Sherief Reda and Abdullah N Nowroz. Power modeling and characterization of computing devices: A survey. Foundations and Trends in Electronic Design Automation, 6(2):121–216, 2012.

[28] Miyuru Dayarathna, Yonggang Wen, and Rui Fan. Data center energy consumption modeling: A survey. IEEE Communications Surveys and Tutorials, 18(1):732–794, jan 2016.

[29] Yogendra Joshi and Pramod Kumar. Introduction to data center energy flow and thermal management, aug 2013.

[30] Lizhe Wang and Samee U Khan. Review of performance metrics for green data centers: a taxonomy study. The journal of supercomputing, 63(3):639–656, 2013.

[31] Anton Beloglazov, Rajkumar Buyya, Young Choon Lee, and Albert Zomaya. A taxonomy and survey of energy-efficient data centers and cloud computing systems. In Advances in computers, volume 82, pages 47–111. Elsevier, 2011.

[32] Sardar Khaliq Uzaman, Atta Ur Rehman Khan, Junaid Shuja, Tahir Maqsood, Faisal Rehman, and Saad Mustafa. A systems overview of commercial data centers: Initial energy and cost analysis. International Journal of Information Technology and Web Engineering, 14(1):42–65, 2019.

[33] Pei Huang, Benedetta Copertaro, Xingxing Zhang, Jingchun Shen, Isabelle Löfgren, Mats Rönnelid, Jan Fahlen, Dan Andersson, and Mikael Svanfeldt. A review of data centers as prosumers in district energy systems: Renewable energy integration and waste heat reuse for district heating, jan 2020.

[34] Sparsh Mittal. A survey of techniques for improving energy efficiency in embedded computing systems. International Journal of Computer Aided Engineering and Technology, 6(4):440–459, 2014.

[35] Tom Bostoen, Sape Mullender, and Yolande Berbers. Power-reduction techniques for data-center storage systems. ACM Computing Surveys (CSUR), 45(3):1–38, 2013.

[36] Jun Wang, Ling Feng, and Wenwei Xue. A review of energy efficiency technology in computer servers and cluster systems. In 2011 3rd International Conference on Computer Research and Development, volume 2, pages 109–113. IEEE, 2011.

[37] Ali Hammadi and Lotfi Mhamdi. A survey on architectures and energy efficiency in data center networks. Computer Communications, 40:1–21, 2014.

[38] Giuseppe Procaccianti and Aristeidis Routsis. Energy efficiency and power measurements: an industrial survey. In ICT for Sustainability 2016, pages 69–78. Atlantis Press, 2016.

[39] Sparsh Mittal. Power management techniques for data centers: A survey. arXiv preprint arXiv:1404.6681, 2014.

[40] Anne-Cecile Orgerie, Marcos Dias de Assuncao, and Laurent Lefevre. A survey on techniques for improving the energy efficiency of large-scale distributed systems. ACM Computing Surveys (CSUR), 46(4):1–31, 2014.

[41] Junaid Shuja, Kashif Bilal, Sajjad A Madani, Mazliza Othman, Rajiv Ranjan, Pavan Balaji, and Samee U Khan. Survey of techniques and architectures for designing energy-efficient data centers. IEEE Systems Journal, 10(2):507–519, 2014.

[42] E. Baccarelli, N. Cordeschi, A. Mei, M. Panella, M. Shojafar, and J. Stefa. Energy-efficient dynamic traffic offloading and reconfiguration of networked data centers for big data stream mobile computing: review, challenges, and a case study. IEEE Network, 30(2):54–61, 2016.

[43] E. S. Madhan and S. Srinivasan. Energy aware data center using dynamic consolidation techniques: A survey. In Proceedings of ICCCS 2014 - IEEE International Conference on Computer Communication and Systems, pages 43–45. Institute of Electrical and Electronics Engineers Inc., mar 2014.

[44] Srimoyee Bhattacherjee, Sunirmal Khatua, and Sarbani Roy. A review on energy efficient resource management strategies for cloud. In Advanced Computing and Systems for Security, pages 3–15. Springer, 2017.

[45] Saleh Atiewi, Salman Yussof, Mohd Ezanee, and Muder Almiani. A review energy-efficient task scheduling algorithms in cloud computing. In 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), pages 1–6. IEEE, 2016.

[46] Christoph Möbius, Waltenegus Dargie, and Alexander Schill. Power consumption estimation models for processors, virtual machines, and servers. IEEE Transactions on Parallel and Distributed Systems, 25(6):1600–1614, 2014.

[47] Nasrin Akhter and Mohamed Othman. Energy aware resource allocation of cloud data center: review and open issues. Cluster Computing, 19(3):1163–1182, sep 2016.

[48] Loiy Alsbatin, Gurcu Oz, and Ali Hakan Ulusoy. An Overview of Energy-Efficient Cloud Data Centres. In 2017 International Conference on Computer and Applications, ICCA 2017, pages 211–214. Institute of Electrical and Electronics Engineers Inc., oct 2017.

[49] Junaid Shuja, Abdullah Gani, Shahaboddin Shamshirband, Raja Wasim Ahmad, and Kashif Bilal. Sustainable Cloud Data Centers: A survey of enabling techniques and technologies. Renewable and Sustainable Energy Reviews, 62:195–214, sep 2016.

[50] Yogendra Joshi and Emad Samadiani. Energy efficient thermal management of data centers via open multi-scale design: A review of research questions and approaches. Journal of Enhanced Heat Transfer, 18(1):15–30, 2011.

[51] Jean Marc Pierson. Large-Scale Distributed Systems and Energy Efficiency. John Wiley & Sons, Inc, Hoboken, NJ, USA, mar 2015.

[52] Kazi Main Uddin Ahmed, Jil Sutaria, Math H.J. Bollen, and Sarah K. Rönnberg. Electrical Energy Consumption Model of Internal Components in Data Centers. Proceedings of 2019 IEEE PES Innovative Smart Grid Technologies Europe, ISGT-Europe 2019, 2019.

[53] Sang Woo Ham, Min Hwi Kim, Byung Nam Choi, and Jae Weon Jeong. Simplified server model to simulate data center cooling energy consumption. Energy and Buildings, 86:328–339, jan 2015.

[54] Frederick K. Frantz. Taxonomy of model abstraction techniques. In Winter Simulation Conference Proceedings, pages 1413–1420, New York, New York, USA, 1995. IEEE.

[55] Suzanne Rivoire, Mehul A. Shah, Parthasarathy Ranganathan, Christos Kozyrakis, and Justin Meza. Models and metrics to enable energy-efficiency optimizations. Computer, 2007.

[56] Micha Vor Dem Berge, Georges Da Costa, Andreas Kopecki, Ariel Oleksiak, Jean Marc Pierson, Tomasz Piontek, Eugen Volk, and Stefan Wesner. Modeling and simulation of data center energy-efficiency in CoolEmAll. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 7396 LNCS, pages 25–36. Springer, Berlin, Heidelberg, 2012.

[57] Avrilia Floratou, Frank Bertsch, Jignesh M. Patel, and Georgios Laskaris. Towards building wind tunnels for data center design. Proceedings of the VLDB Endowment, 7(9):781–784, may 2014.

[58] Xiao Zhang, Jian Jun Lu, Xiao Qin, and Xiao Nan Zhao. A high-level energy consumption model for heterogeneous data centers. Simulation Modelling Practice and Theory, 39:41–55, dec 2013.

[59] Daniel C. Kilper, Gary Atkinson, Steven K. Korotky, Suresh Goyal, Peter Vetter, Dusan Suvakovic, and Oliver Blume. Power trends in communication networks. IEEE Journal on Selected Topics in Quantum Electronics, 17(2):275–284, mar 2011.

[60] Yichao Jin, Yonggang Wen, Qinghua Chen, and Zuqing Zhu. An empirical investigation of the impact of server virtualization on energy efficiency for green data center. The Computer Journal, 56(8):977–990, aug 2013.

[61] Morgan Tatchell-Evans, Nik Kapur, Jonathan Summers, Harvey Thompson, and Dan Oldham. An experimental and theoretical investigation of the extent of bypass air within data centres employing aisle containment, and its impact on power consumption. Applied Energy, 2017.

[62] Swapnoneel Roy, Atri Rudra, and Akshat Verma. An energy complexity model for algorithms. In ITCS 2013 - Proceedings of the 2013 ACM Conference on Innovations in Theoretical Computer Science, pages 283–303, New York, New York, USA, 2013. ACM Press.

[63] Bogdan Marius Tudor and Yong Meng Teo. On understanding the energy consumption of ARM-based multicore servers. In Proceedings of

the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems - SIGMETRICS '13, page 267, New York, New York, USA, 2013. Association for Computing Machinery (ACM).

[64] Shuaiwen Leon Song, Kevin Barker, and Darren Kerbyson. Unified performance and power modeling of scientific workloads. In Proc. of E2SC 2013: 1st Int. Workshop on Energy Efficient Supercomputing - Held in Conjunction with SC 2013: The Int. Conference for High Performance Computing, Networking, Storage and Analysis, 2013.

[65] Power-conservative server consolidation based resource management in cloud. International Journal of Network Management, 24(6):415–432, nov 2014.

[66] Angelos Chatzipapas, Dimosthenis Pediaditakis, Charalampos Rotsos, Vincenzo Mancuso, Jon Crowcroft, and Andrew W. Moore. Challenge: Resolving data center power bill disputes: The energy-performance trade-offs of consolidation. In e-Energy 2015 - Proceedings of the 2015 ACM 6th International Conference on Future Energy Systems, pages 89–94, New York, New York, USA, jul 2015. Association for Computing Machinery, Inc.

[67] Ismail Alan, Engin Arslan, and Tevfik Kosar. Energy-Aware Data Transfer Tuning. 2014.

[68] Adam Lewis, Soumik Ghosh, and N.-F. Tzeng. Run-time energy consumption estimation based on workload in server systems. Proceedings of the 2008 conference on Power aware computing and systems, pages 1–4, 2008.

[69] W. Lloyd Bircher and Lizy Kurian John. Core-level activity prediction for multicore power management. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 1(3):218–227, sep 2011.

[70] Rong Ge, Xizhou Feng, and Kirk W. Cameron. Modeling and evaluating energy-performance efficiency of parallel processing on multicore based power aware systems. In IPDPS 2009 - Proceedings of the 2009 IEEE International Parallel and Distributed Processing Symposium, 2009.

[71] Runtime energy consumption estimation for server workloads based on chaotic time-series approximation. In Transactions on Architecture and Code Optimization, volume 9, pages 1–26, sep 2012.

[72] Robert Basmadjian, Nasir Ali, Florian Niedermeier, and Giovanni Giuliani. A Methodology to Predict the Power Consumption of Servers in Data Centres. In e-Energy '11: Proceedings of the 2nd International Conference on Energy-Efficient Computing and Networking. Association for Computing Machinery, New York, NY, United States, 2011.

[73] Gaurav Dhiman, Kresimir Mihic, and Tajana Rosing. A system for online power prediction in virtualized environments using Gaussian mixture models. Proceedings - Design Automation Conference, (3):807–812, 2010.

[74] Peng Xiao, Zhigang Hu, Dongbo Liu, Guofeng Yan, and Xilong Qu. Virtual machine power measuring technique with bounded error in cloud environments. Journal of Network and Computer Applications, 36(2):818–828, 2013.

[75] Feifei Chen, John Grundy, Yun Yang, Jean Guy Schneider, and Qiang He. Experimental analysis of task-based energy consumption in cloud computing systems, 2013.

[76] Peter Garraghan, Harvey Thompson, Yaser Al-Anii, Nik Kapur, Jon Summers, and Karim Djemame. A unified model for holistic power usage in cloud datacenter servers. Proceedings - 9th IEEE/ACM International Conference on Utility and Cloud Computing, UCC 2016, pages 11–19, 2016.

[77] A Power Consumption Model for Cloud Servers Based on Elman Neural Network. IEEE Transactions on Cloud Computing, pages 1–1, jun 2019.

[78] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso. Power provisioning for a warehouse-sized computer. ACM SIGARCH Computer Architecture News, 35(2):13, jun 2007.

[79] Frank Bellosa. The benefits of event-driven energy accounting in power-sensitive systems. Technical report, 2000.

[80] Richard Kavanagh and Karim Djemame. Rapid and accurate energy models through calibration with IPMI and RAPL. Concurrency Computation, 31(13):1–21, 2019.

[81] Saif Ul Islam and Jean Marc Pierson. Evaluating energy consumption in CDN servers. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 7453 LNCS, pages 64–78. Springer, Berlin, Heidelberg, 2012.

[82] Vishal Gupta, Ripal Nathuji, and Karsten Schwan. An analysis of power reduction in datacenters using heterogeneous chip multiprocessors. Technical Report 3, 2011.

[83] Anton Beloglazov, Jemal Abawajy, and Rajkumar Buyya. Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. Future Generation Computer Systems, 28(5):755–768, may 2012.

[84] Dimitris Economou, Suzanne Rivoire, Christos Kozyrakis, and Partha Ranganathan. Full-System Power Analysis and Modeling for Server Environments. Technical Report 3, 2006.

[85] Ricardo Lent. A model for network server performance and power consumption. Sustainable Computing: Informatics and Systems, 3(2):80–93, jun 2013.

[86] Yanfei Li, Ying Wang, Bo Yin, and Lu Guan. An online power metering model for cloud environment. In Proceedings - IEEE 11th International Symposium on Network Computing and Applications, NCA 2012, pages 175–180, 2012.

[87] Aman Kansal, Feng Zhao, Jie Liu, Nupur Kothari, and Arka A. Bhattacharya. Virtual machine power metering and provisioning. 2010.

[88] Georges Da Costa and Helmut Hlavacs. Methodology of measurement for energy consumption of applications. In Proceedings - IEEE/ACM International Workshop on Grid Computing, pages 290–297, 2010.

[89] M. Witkowski, A. Oleksiak, T. Piontek, and J. Weoglarz. Practical power consumption estimation for real life HPC applications. Future Generation Computer Systems, 29(1):208–217, jan 2013.

[90] Adam Lewis, Soumik Ghosh, and N.-F. Tzeng. Run-time energy consumption estimation based on workload in server systems. Technical report, 2008.

[91] Suzanne Rivoire, Parthasarathy Ranganathan, and Christos Kozyrakis. A comparison of high-level full-system power models. In Proceedings of the 2008 Conference on Power Aware Computing and Systems, HotPower'08, page 3, USA, 2008. USENIX Association.

[92] Weiwei Lin, Weiqi Wang, Wentai Wu, Xiongwen Pang, Bo Liu, and Ying Zhang. A heuristic task scheduling algorithm based on server power efficiency model in cloud environments. Sustainable Computing: Informatics and Systems, 20:56–65, dec 2018.

[93] E. N. Mootaz Elnozahy, Michael Kistler, and Ramakrishnan Rajamony. Energy-efficient server clusters. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 2325, pages 179–197. Springer Verlag, 2003.

[94] Hui Li, Giuliano Casale, and Tariq Ellahi. SLA-driven planning and optimization of enterprise applications. 2010.

[95] Yongqiang Gao, Haibing Guan, Zhengwei Qi, Bin Wang, and Liang Liu. Quality of service aware power management for virtualized data centers. Journal of Systems Architecture, 59(4-5):245–259, apr 2013.

[96] Maolin Tang and Shenchen Pan. A Hybrid Genetic Algorithm for the Energy-Efficient Virtual Machine Placement Problem in Data Centers. Neural Processing Letters, 41(2):211–221, jan 2015.

[97] Corey Gough, Ian Steiner, and Winston Saunders. Energy efficient servers: Blueprints for data center optimization. Apress Media LLC, jan 2015.

[98] Dynamo: Facebook's Data Center-Wide Power Management System. In Proceedings - 2016 43rd International Symposium on Computer Architecture, ISCA 2016, 2016.

[99] Wei Huang, Malcolm Allen-Ware, John B. Carter, Elmootazbellah Elnozahy, Hendrik Hamann, Tom Keller, Charles Lefurgy, Jian Li, Karthick Rajamani, and Juan Rubio. TAPO: Thermal-aware power optimization techniques for servers and data centers. In 2011 International Green Computing Conference and Workshops, IGCC 2011, 2011.

[100] Donghwa Shin, Jihun Kim, Naehyuck Chang, Jinhang Choi, Sung Woo Chung, and Eui Young Chung. Energy-optimal dynamic thermal management for green computing. In IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD, 2009.

[101] Kazi Main Uddin Ahmed, Roger Alves de Oliveira, Manuel Alvarez, and Math H. J. Bollen. Risk Assessment of Server Outages Due To Voltage Dips In the Internal Power Supply System of a Data Center. In 26th International Conference on Electricity Distribution, Online, 20-23 September 2021. IET.

[102] Kaiyu Sun, Na Luo, Xuan Luo, and Tianzhen Hong. Prototype energy models for data centers. Energy and Buildings, 231:110603, jan 2021.

[103] Soha Rawas. Energy, network, and application-aware virtual machine placement model in SDN-enabled large scale cloud data centers. Multimedia Tools and Applications, pages 1–22, feb 2021.

[104] Yuanxiong Guo and Yuguang Fang. Electricity cost saving strategy in data centers by using energy storage. IEEE Transactions on Parallel and Distributed Systems, 24(6):1149–1160, 2013.

[105] Ruben Milocco, Pascale Minet, Eric Renault, and Selma Boumerdassi. Proactive Data Center Management Using Predictive Approaches. IEEE Access, 8:161776–161786, sep 2020.

[106] Rajesh Bose, Sandip Roy, Haraprasad Mondal, Dipan Roy Chowdhury, and Srabanti Chakraborty. Energy-efficient approach to lower the carbon emissions of data centers. Computing, pages 1–19, jan 2021.

[107] Bo Li, Peng Hou, Hao Wu, Rongrong Qian, and Hongwei Ding. Placement of edge server based on task overhead in mobile edge computing environment. Transactions on Emerging Telecommunications Technologies, 2020.

[108] Yuanzhe Li and Shangguang Wang. An energy-aware edge server placement algorithm in mobile edge computing. In Proceedings - 2018 IEEE International Conference on Edge Computing, EDGE 2018 - Part of the 2018 IEEE World Congress on Services, pages 66–73. Institute of Electrical and Electronics Engineers Inc., sep 2018.

[109] A proactive autoscaling and energy-efficient VM allocation framework using online multi-resource neural network for cloud data center. Neurocomputing, 426:248–264, feb 2021.

[110] Rahul Yadav, Weizhe Zhang, Keqin Li, Chuanyi Liu, and Asif Ali Laghari. Managing overloaded hosts for energy-efficiency in cloud data centers. Cluster Computing, pages 1–15, feb 2021.

[111] Steven Pelley, David Mesiner, Pooya Zandevakili, Thomas F. Wenisch, and Jack Underwood. Power routing: Dynamic power provisioning in the data center. ACM SIGPLAN Notices, 45(3):231–242, mar 2010.

[112] Neil Rasmussen. Calculating Total Cooling Requirements for Data Centers. Technical report, 2011.

[113] Kazi Main Uddin Ahmed, Manuel Alvarez, and Math H. J. Bollen. A novel reliability index to assess the computational resource adequacy in data centers. IEEE Access, 9:54530–54541, 2021.

[114] Annabelle Pratt, Pavan Kumar, and Tomm V. Aldridge. Evaluation of 400V DC distribution in telco and data centers to improve energy efficiency. In INTELEC, International Telecommunications Energy Conference (Proceedings), pages 32–39, 2007.

[115] Steven Pelley, David Meisner, Thomas F Wenisch, and James W VanGilder. Understanding and Abstracting Total Data Center Power. Technical report, 2009.

[116] Mohammed Faham Alsolami, Karun Arjun Potty, and Jin Wang. Mitigation of double-line-frequency current ripple in switched capacitor based ups system. IEEE Transactions on Power Electronics, 36(4):4042–4051, 2020.

[117] Ataollah Gogani Khiabani and Ali Heydari. Design and implementation of an optimal switching controller for uninterruptible power supply inverters using adaptive dynamic programming. IET Power Electronics, 12(12):3068–3076, 2019.

[118] Leonardo Göbel Fernandes, Alceu André Badin, Daniel Flores Cortez, Roger Gules, Eduardo Félix Ribeiro Romaneli, and Amauri Assef. Transformerless ups system based on the half-bridge hybrid switched-capacitor operating as ac–dc and dc–dc converter. IEEE Transactions on Industrial Electronics, 68(3):2173–2183, 2020.

[119] Kaixia Lu and Li Huang. Daily maintenance and fault handling of ups signal power supply system in wuhan metro. In MIPPR 2019: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications, volume 11432, page 114321E. International Society for Optics and Photonics, 2020.

[120] Qiongbin Lin, Fenghuang Cai, Wu Wang, Sixiong Chen, Zhe Zhang, and Shi You. A high-performance online uninterruptible power supply (ups) system based on multitask decomposition. IEEE Transactions on Industry Applications, 55(6):7575–7585, 2019.

[121] Neil Rasmussen. Electrical Efficiency Modeling for Data Centers_113. Technical Report 0-113, 2007.

[122] Ehsan Pakbaznia and Massoud Pedram. Minimizing data center cooling and server power costs. In Proceedings of the International Symposium on Low Power Electronics and Design, pages 145–150, New York, New York, USA, 2009. ACM Press.

[123] Zahra Abbasi, Georgios Varsamopoulos, and Sandeep K.S. Gupta. TACOMA: Server and workload management in internet data centers considering cooling-computing power trade-off and energy proportionality. Transactions on Architecture and Code Optimization, 9(2):1–37, jun 2012.

[124] Eun Kyung Lee, Indraneel Kulkarni, Dario Pompili, and Manish Parashar. Proactive thermal management in green datacenters. Journal of Supercomputing, 60(2):165–195, may 2012.

[125] Reza Azimi, Xin Zhan, and Sherief Reda. Thermal-aware layout planning for heterogeneous datacenters. In Proceedings of the International Symposium on Low Power Electronics and Design, volume 2015-October, pages 245–250. Institute of Electrical and Electronics Engineers Inc., oct 2015.

[126] Sahar Asgari, Seyed Morteza MirhoseiniNejad, Hosein Moazamigoodarzi, Rohit Gupta, Rong Zheng, and Ishwar K. Puri. A gray-box model for real-time transient temperature predictions in data centers. Applied Thermal Engineering, 185:116319, feb 2021.

[127] Chisato Matsuda and Yosuke Mino. Study on power-saving effects in direct-use of geothermal energy for datacenter cooling systems. In INTELEC, International Telecommunications Energy Conference (Proceedings), volume 2016-November. Institute of Electrical and Electronics Engineers Inc., nov 2016.

[128] H. F. Hamann, T. G. Van Kessel, M. Iyengar, J. Y. Chung, W. Hirt, M. A. Schappert, A. Claassen, J. M. Cook, W. Min, Y. Amemiya, V. López, J. A. Lacey, and M. O'Boyle. Uncovering energy-efficiency opportunities in data centers. IBM Journal of Research and Development, 53(3), 2009.

[129] Satoshi Itoh, Yuetsu Kodama, Toshiyuki Shimizu, Satoshi Sekiguchi, Hiroshi Nakamura, and Naohiko Mori. Power consumption and efficiency of cooling in a data center. In Grid Computing (GRID), 2010 11th IEEE/ACM International Conference on, pages 305–312. IEEE, 2010.

[130] Kosuke Sasakura, Takeshi Aoki, Masayoshi Komatsu, and Takeshi Watanabe. Rack temperature prediction model using machine learning after stopping computer room air conditioner in server room. Energies, 13(17):4300, 2020.

[131] Hao Tian, Hang Liang, and Zhen Li. A new mathematical model for multi-scale thermal management of data centers using entransy theory. Building Simulation, 12(2):323–336, apr 2019.

[132] Rasoul Rahmani, Irene Moser, and Mohammadmehdi Seyedmahmoudian. A complete model for modular simulation of data centre power load. arXiv preprint arXiv:1804.00703, 2018.

[133] Rajarshi Das, Jeffrey O. Kephart, Jonathan Lenchner, and Hendrik Hamann. Utility-function-driven energy-efficient cooling in data centers. In Proceeding of the 7th International Conference on Autonomic Computing, ICAC '10 and Co-located Workshops, pages 61–70, New York, New York, USA, 2010. ACM Press.

[134] J. Endrenyi. Reliability modeling in electric power systems. Wiley, 1978.

[135] Roy Billinton, Ronald N. Allan, Roy Billinton, and Ronald N. Allan. Reliability Evaluation of Engineering Systems. In Reliability Evaluation of Engineering Systems, pages 1–20. Springer US, 1992.

[136] Robert Arno, Addam Friedl, Peter Gross, and Robert J. Schuerger. Reliability of data centers by tier classification. In IEEE Transactions on Industry Applications, volume 48, pages 777–783, mar 2012.

[137] Pitt Turner, John H Seader, Vince Renaud, and Kenneth G Brill. Tier Classifications Define Site Infrastructure Performance. Uptime Institute, page 20, 2008.

[138] W Pitt, Turner Iv, John H Seader, and Kenneth G Brill. Industry Standard Tier Classifi cations Defi ne Site Infrastructure Performance The Uptime Institute Industry Standard Tier Classifi cations Defi ne Site Infrastructure Performance. Santa Fe, NM: Uptime Institute, 2001.

[139] Hwaiyu Geng. Data Center Handbook. Wiley, oct 2014.

[140] B Erkus, SS Polat, and H Darama. Seismic design of data centers for tier iii and tier iv resilience: Basis of design. In 11^th US National Conference on Earthquake Engineering. Integrating Science, Engineering & Policy.

[141] IEEE. Design of Reliable Industrial and Commercial Power Systems, volume 2007. 2007.

[142] Eric Bauer, Randee Adams, and Daniel Eustace. Beyond redundancy: how geographic redundancy can improve service availability and reliability of computer-based systems. John Wiley & Sons, 2011.

[143] S. U. Amin and M. S. Hossain. Edge intelligence and internet of things in healthcare: A survey. IEEE Access, 9:45–59, 2021.

[144] Polona Štefanič, Petar Kochovski, Omer F Rana, and Vlado Stankovski. Quality of service-aware matchmaking for adaptive microservice-based applications. Concurrency and Computation: Practice and Experience, page e6120, 2020.

[145] Chee Wei Ang and Chen Khong Tham. Analysis and optimization of service availability in an HA cluster with load-dependent machine availability. IEEE Transactions on Parallel and Distributed Systems, 18(9):1307–1319, 2007.

[146] Kiran Nagaraja, Gustavo Gama, Ricardo Bianchini, Richard P. Martin, Wagner Meira, and Thu D. Nguyen. Quantifying the performability of cluster-based services. IEEE Transactions on Parallel and Distributed Systems, 16(5):456–467, may 2005.

[147] Michael K. Patterson. The effect of data center temperature on energy efficiency. 2008 11th IEEE Intersociety Conference on Thermal and

Thermomechanical Phenomena in Electronic Systems, I-THERM, pages 1167–1174, 2008.

[148] Walid Mokhtar Bennaceur and Leïla Kloul. Formal models for safety and performance analysis of a data center system. Reliability Engineering and System Safety, 193:106643, jan 2020.

[149] Xiao-Yang Li, Yue Liu, Yan-Hui Lin, Liang-Hua Xiao, Enrico Zio, and Rui Kang. A generalized petri net-based modeling framework for service reliability evaluation and management of cloud data centers. Reliability Engineering & System Safety, 207:107381, 2021.

[150] Misbah Liaqat, Anjum Naveed, Rana Liaqat Ali, Junaid Shuja, and Kwang-Man Ko. Characterizing dynamic load balancing in cloud environments using virtual machine deployment models. IEEE Access, 7:145767–145776, 2019.

[151] Saad Mustafa, Kinza Sattar, Junaid Shuja, Shahzad Sarwar, Tahir Maqsood, Sajjad A. Madani, and Sghaier Guizani. Sla-aware best fit decreasing techniques for workload consolidation in clouds. IEEE Access, 7:135256–135267, 2019.

[152] Jiaqiang Wang, Quan Zhang, Sungmin Yoon, and Yuebin Yu. Reliability and availability analysis of a hybrid cooling system with water-side economizer in data center. Building and Environment, 148:405–416, jan 2019.

[153] Ryuichi Nishida, Shisei Waragai, Keisuke Sekiguchi, Manabu Kishita, Hiroaki Miyake, and Tsuneo Uekusa. Relationship between the reliability of a data-center air-conditioning system and the air-conditioning power supply. In INTELEC, International Telecommunications Energy Conference (Proceedings), 2008.

[154] Howard Cheung and Shengwei Wang. Reliability and availability assessment and enhancement of water-cooled multi-chiller cooling systems for data centers. Reliability Engineering and System Safety, 191:106573, nov 2019.

[155] Jiang Jiang Wang, Chao Fu, Kun Yang, Xu Tao Zhang, Guo hua Shi, and John Zhai. Reliability and availability analysis of redundant BCHP (building cooling, heating and power) system. Energy, 61:531–540, nov 2013.

[156] Yang Lei and Alex Q. Huang. Data center power distribution system reliability analysis tool based on Monte Carlo next event simulation method. In 2017 IEEE Energy Conversion Congress and Exposition, ECCE 2017, volume 2017-January, pages 2031–2035. Institute of Electrical and Electronics Engineers Inc., nov 2017.

[157] Katsuichi Yotsumoto, Seiichi Muroyama, Shoji Matsumura, and Hitoshi Watanabe. Design for a highly efficient distributed power supply system based on reliability analysis. In INTELEC, International Telecommunications Energy Conference (Proceedings), pages 545–550. Publ by IEEE, 1988.

[158] V. Sithimolada and P. W. Sauer. Facility-level DC vs. typical AC distribution for data centers: A comparative reliability study. IEEE Region 10 Annual International Conference, Proceedings/TENCON, pages 2102–2107, 2010.

[159] Daniel Rosendo, Demis Gomes, Guto Leoni Santos, Glauco Goncalves, Andre Moreira, Leylane Ferreira, Patricia Takako Endo, Judith Kelner, Djamel Sadok, Amardeep Mehta, and Mattias Wildeman. A methodology to assess the availability of next-generation data centers. Journal of Supercomputing, 75(10):6361–6385, 2019.

[160] Bijen Raj Shrestha, Timothy M. Hansen, and Reinaldo Tonkoski. Reliability analysis of 380V DC distribution in data centers. 2016 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2016, pages 0–4, 2016.

[161] Alexander Barthelme, Xiwen Xu, and Tiefu Zhao. A hybrid AC and DC distribution architecture in data centers. 2017 IEEE Energy Conversion Congress and Exposition, ECCE 2017, 2017-January:2017–2022, 2017.

[162] Wei-Jenn Ke and Sheng-De Wang. Reliability evaluation for distributed computing networks with imperfect nodes. IEEE Transactions on Reliability, 46(3):342–349, 1997.

[163] Mini S Thomas and Ikbal Ali. Reliable, fast, and deterministic substation communication network architecture and its performance simulation. IEEE Transactions on Power Delivery, 25(4):2364–2370, 2010.

[164] Jiajia Chen, Lena Wosinska, Mohsan Niaz Chughtai, and Marco Forzati. Scalable passive optical network architecture for reliable service delivery. Journal of Optical Communications and Networking, 3(9):667–673, 2011.

[165] Rodrigo de Souza Couto, Stefano Secci, Miguel Elias Mitre Campista, and Luís Henrique Maciel Kosmalski Costa. Reliability and Survivability Analysis of Data Center Network Topologies. Journal of Network and Systems Management, 24(2):346–392, apr 2016.

[166] Carlos Colman-Meixner, Chris Develder, Massimo Tornatore, and Biswanath Mukherjee. A survey on resiliency techniques in cloud computing infrastructures and applications. IEEE Communications Surveys & Tutorials, 18(3):2244–2281, 2016.

[167] D. J. Rosenkrantz, S. Goel, S. S. Ravi, and J. Gangolly. Resilience metrics for service-oriented networks: A service allocation approach. IEEE Transactions on Services Computing, 2(3):183–196, 2009.

[168] Gustavo Callou, João Ferreira, Paulo Maciel, Dietmar Tutsch, and Rafael Souza. An Integrated Modeling Approach to Evaluate and Optimize Data Center Sustainability, Dependability and Cost. mdpi.com, 7:238–277, 2014.

[169] Ryan Robidoux, Haiping Xu, Liudong Xing, and Mengchu Zhou. Automated modeling of dynamic reliability block diagrams using colored Petri nets. IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans, 40(2):337–351, mar 2010.

[170] Mohd Khairil Rahmat and Muhammad Nadzmi Sani. Fault tree analysis in UPS reliability estimation. In 2014 4th International Conference on Engineering Technology and Technopreneuship, ICE2T 2014, volume 2014-August, pages 240–245. Institute of Electrical and Electronics Engineers Inc., jan 2015.

[171] Thanyalak Chalermarrewong, Tiranee Achalakul, and Simon Chong Wee See. Failure prediction of data centers using time series and fault tree analysis. In 2012 IEEE 18th International Conference on Parallel and Distributed Systems, pages 794–799, 2012.

[172] Getzi Jeba Leelipushpam, Immanuel Johnraja Jebadurai, and Jebaveerasingh Jebadurai. Fault tree analysis based virtual machine migration for fault-tolerant cloud data center. Journal of Integrated Design and Process Science, (Preprint):1–17, 2019.

[173] C Heising. IEEE recommended practice for the design of reliable industrial and commercial power systems. IEEE Inc., New York, 2007.

[174] Math H.J. Bollen. Literature search for reliability data of components in electric Literature Search for Reliability Data of Components in Electric Distribution Networks. Technical Report 1993, Electrical Engineering, Technische Universiteit Eindhoven, 1993.

[175] Joseph Hellerstein Charles Reiss, John Wilkes. Google cluster-usage traces format schema 2014-11-17 external.pdf - Google Drive. Technical report, Google Inc., 2014.

[176] Bianca Schroeder and Garth Gibson. A large-scale study of failures in high-performance computing systems. IEEE transactions on Dependable and Secure Computing, 7(4):337–350, 2009.

[177] Tessema M. Mengistu, Dunren Che, Abdulrahman Alahmadi, and Shiyong Lu. Semi-markov process based reliability and availability prediction for volunteer cloud systems. In 2018 IEEE 11th International Conference on Cloud Computing (CLOUD), pages 359–366, 2018.

[178] S. Jeyalaksshmi, M.S. Nidhya, G. Suseendran, Souvik Pal, and D. Akila. Developing mapping and allotment in volunteer cloud systems using reliability profile algorithms in a virtual machine. In 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), pages 97–101, 2021.

[179] Montri Wiboonrat. An empirical study on data center system failure diagnosis. In 2008 The Third International Conference on Internet Monitoring and Protection, pages 103–108, 2008.

[180] MichaÅĆ Aibin and Krzysztof Walkowiak. Monte carlo tree search for cross-stratum optimization of survivable inter-data center elastic optical network. In 2018 10th International Workshop on Resilient Networks Design and Modeling (RNDM), pages 1–7, 2018.

[181] Wang Gang, Ding Mao-Sheng, and Li Xiao-Hua. Analysis of ups system reliability based on monte carlo approach. In 2004 IEEE Region 10 Conference TENCON 2004., volume D, pages 205–208 Vol. 4, 2004.

[182] Mohd Khairil Rahmat, Slobodan Jovanovic, and Kwok Lun Lo. Reliability and availability modelling of uninterruptible power supply (ups) systems using monte-carlo simulation. In 2011 5th International Power Engineering and Optimization Conference, pages 267–272, 2011.

[183] Kazi Main Uddin Ahmed, Manuel Alvarez, and Math H.J. Bollen. Characterizing failure and repair time of servers in a hyper-scale data center. In IEEE PES Innovative Smart Grid Technologies Conference Europe, volume 2020-October, pages 660–664. IEEE, oct 2020.

[184] Mohammad Reza Mesbahi, Amir Masoud Rahmani, and Mehdi Hosseinzadeh. Cloud dependability analysis: Characterizing google cluster infrastructure reliability. In 2017 3th International Conference on Web Research (ICWR), pages 56–61. IEEE, 2017.

[185] Mohammad Mesbahi, Amir Rahmani, and Mehdi Hosseinzadeh. Highly reliable architecture using the 80/20 rule in cloud computing datacenters. Future Generation Computer Systems, 77(C):77–86, 2017.

[186] W. Earl Smith, Kishor S. Trivedi, Lorrie A. Tomek, and Jerry Ackaret. Availability analysis of blade server systems. IBM Systems Journal, 47(4):621–640, 2008.
[187] William S Griffith. Optimal reliability modeling: principles and applications, 2004.
[188] Algirdas Avižienis, Jean Claude Laprie, Brian Randell, and Carl Landwehr. Basic concepts and taxonomy of dependable and secure computing. IEEE Transactions on Dependable and Secure Computing, 1(1):11–33, jan 2004.
[189] Fernando Paganini, Diego Goldsztajn, and Andres Ferragut. An optimization approach to load balancing, scheduling and right sizing of cloud computing systems with data locality. In Proceedings of the IEEE Conference on Decision and Control, volume 2019-December, pages 1114–1119. Institute of Electrical and Electronics Engineers Inc., dec 2019.
[190] Susanne Albers and Jens Quedenfeld. Optimal algorithms for right-sizing data centers. In Annual ACM Symposium on Parallelism in Algorithms and Architectures, pages 363–372, New York, NY, USA, jul 2018. Association for Computing Machinery.
[191] Jeffrey S. Chase, Darrell C. Anderson, Prachi N. Thakar, Amin M. Vahdat, and Ronald P. Doyle. Managing energy and server resources in hosting centers. Operating Systems Review (ACM), 35(5):103–116, dec 2001.
[192] Dian Shen, Junzhou Luo, Fang Dong, Xiang Fei, Wei Wang, Guoqing Jin, and Weidong Li. Stochastic modeling of dynamic right-sizing for energy-efficiency in cloud data centers. Future Generation Computer Systems, 48:82–95, jul 2015.
[193] Minghong Lin, Adam Wierman, Lachlan L H Andrew, Senior Member, and Eno Thereska. Dynamic Right-Sizing for Power-Proportional Data Centers. 21(5):1378–1391, 2013.

KAZI MAIN UDDIN AHMED (Student Member, IEEE) received the M.Sc. degree in electrical engineering jointly from the KTH Royal Institute of Technology, Stockholm, Sweden and the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2014. He is currently pursuing the Ph.D. degree with the Electric Power Engineering Group, Luleå University of Technology, Skellefteå, Sweden.

He was a Lecturer of power system subjects in department of electrical engineering of University of Asia Pacific and North South University, Bangladesh from 2014 to 2018. His research interests include industrial power systems, power systems analysis, energy management, and renewable energy aspects in smart grid.

MANUEL ALVAREZ (M'19) is currently an Associate Senior Lecturer in the Department of Engineering Sciences and Mathematics with the Luleå University of Technology in Skellefteå, Sweden. He received the E.E. (2006) and M.Sc. (2009) degrees in electric power engineering from the Simón Bolívar University in Caracas, Venezuela. He received both the Tk.L. (2017) and the Ph.D.(2019) degrees from the Luleå University of Technology in Skellefteå, Sweden. His research is oriented towards power systems operation and planning, electricity markets, and integration of renewable energy.

MATH H. J. BOLLEN (M'93-SM'96-F'05) received the MSc and PhD degrees from Eindhoven University of Technology, Eindhoven, The Netherlands, in 1985 and 1989, respectively. Currently, he is professor in electric power engineering at Luleå University of Technology, Skellefteå, Sweden. Earlier he has among others been, lecturer at the University of Manchester Institute of Science and Technology (UMIST), Manchester, U.K., R&D manager and technical manager power quality and distributed generation at STRI AB, Gothenburg, Sweden, and technical expect at the Energy Markets Inspectorate, Eskilstuna, Sweden.

He has published a few hundred papers including a number of fundamental papers on voltage dip analysis, two textbooks on power quality, "understanding power quality problems" and "signal processing of power quality disturbances", and two textbooks on the future power system: "integration of distributed generation in the power system" and "the smart grid - adapting the power system to new challenges".

Nomenclature

$A_f(\infty)$	functional availability of cooling system
$A_o(\infty)$	operational availability of cooling system
P_{idle}	average power consumption of server in idle mode
P_{max}	maximum average power consumption of server when it is fully utilized
η_{heat}	efficiency of the CRAC unit
ϕ_{PDU}	power loss coefficient of PDU
ϕ_{UPS}	power loss coefficient of UPS
C_{CoP}	coefficent of performance of the CRACR unit
C_{cpu}	coefficients of CPU power consumption
C_{disk}	coefficients of hard disk drive power consumption
C_{memory}	coefficients of memory unit power consumption
C_{NIC}	coefficients of NIC power consumption
$Conn_{Max_s}$	maximum number of connections allowed on the server s
$Conn_{s[t_1,t_2]}$	actual number of connections to the server s between time interval t_1 and t_2
E_{board}	energy consumed by peripherals that support the operation of the board
E_{CPU}	energy consumption of the CPU
$E_{CPU}(A)$	energy consumption of a CPU while running the algorithm A
E_{disk}	energy consumption of the disk driver in the server
E_{em}	energy consumed by the electro-mechanical components in the blade server including fans
E_{hdd}	energy consumed by the hard disk drive during the server's operation
$E_{I/O}$	energy consumption by the input/output peripheral slot of the server
E_{mem}	energy consumption of the memory
$E_{mem}(A)$	energy consumption of a memory while running the algorithm A
E_{NIC}	energy consumption of the network interface card in the server
E_{server}	energy consumption of the server
$E_{server}(A)$	energy consumption of a server while running the algorithm A
H_{active}	active state power consumption of the host server
H_{idle}	idle power consumption of the host server
i	total number of mainboards or motherboards
k	total number of PSU attached with the server
M	number of active VMs on this server
m	total number of fans attached with the server
n	total number of pumps attached to the rack
p_t	probability of intended value of room temperature when the cooling system fails
P_Δ	correction factor of the server power consumption model
P_{active}	active state power consumption of the server
P_{base}	base power consumption of the server
P_{comp}	combined CPU and memory average power usage
P_{CPU}	power consumption of the CPU
$P_{CRAC_{cool}}$	power consumption of CRAR unit
P_{disk}	power consumption of the disk drivers
P_{Fan_j}	total power consumption of the local fans
P_{fix}	fixed power consumption of the server and the cooling system
$P_{I/O}$	power consumption by the input/output peripheral slot
P_{mb_i}	total power consumption or conduction loss of the mainboards
P_{mem}	power consumption of the memory units
$P_{net_{dev}}$	power consumption of of the network devices
P_{NIC}	average power consumption of the network interface card
P_{PDU}^{idle}	idle power loss of PDU
P_{PDU}^{Loss}	power loss of PDU
P_{PSU_k}	total power consumption of the PSUs
P_{Pump}	power consumption of the pump
P_r	power consumption of the refrigeration system
P_{server}	server power consumption
P_{sf}^{max}	maximum amount of heat equivalent power that can be generated from the server
P_t	power consumption of server at time t
P_{UPS}^{idle}	idle power loss of UPS
P_{UPS}^{Loss}	power loss of UPS
p_{us}	probability of room temperature out of intended range when the system is under operation state
P_{var}	power consumed by running tasks in the cloud system
P_{VM}	dynamic power consumption of a specific VM
Q_{inlet}	inlet heat of the racks
T_{comm}	total network usage time
T_{comp}	average computation time
T_{die}	die temperature of CPU
$T_{net_{dev}}$	average running time of the netwrok devices
t_s	supplied coolant temperature of the CRAC unit
U_x^{uti}	total CPU utilization of x^{th} host server
U_y^{uti}	CPU utilization by y^{th} VM
U_{count}	total VMs assigned in the host server
u_{cpu_t}	CPU utilization at time t
u_{cpu}	CPU average utilization
u_{disk_t}	hard disk I/O request rate at time t
u_{disk}	average utilization of disk
u_{mem_t}	memory access rate at time t
u_{mem}	average utilization of memory
u_{net_t}	network I/O request rate at time t
u_{net}	average utilization network device
W_i	processor utilization ratio allocated to i_{th} VM