Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2018, 5(6):157-163**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Optimization of Energy Consumption for Data Center

FRANCIS OKOYE[1], TOOCHUKWU ASOGWA[2], GODWIN OZOR[1]

[1]Computer Engineering, [2]Computer Science
Enugu State University of Science and Technology

Abstract Energy consumption optimization is the key to efficient use of the commodity called 'energy' in any given process or plant or center. Data center is an energy demanding infrastructures that runs large scale internet related services. Energy consumption optimization is vital to designing and development of an efficient model that reduces excessive energy demand and even forecast in data centers. This paper investigates the energy requirement of the 'major' components or units or modules of the data center infrastructure both at peak operations and idle state. The simulation results indicate significant energy management without compromising the quality of service.

Keywords Energy consumption, data center, optimization, IOT.

1. Introduction

Data center can be expressed as an energy demanding computing infrastructure. Data services from data centers are configured or designed or known for zero downtime, hence 24hours uplink and downlink to propel the fast growth of IT industry and transform the economy at large [1]. The ever increasing growth in the demand for data computing, processing and storage by a variety of large scale services, such as Google and Facebook, by telecommunication operators such as Nigeria based telecoms, by banks, by oil and gas industries for pipeline monitoring, research institutions and others, resulted in the proliferation of large data centers with thousands of servers. The requirement for supporting a vast variety of applications ranging from those that run for a few seconds to those that run persistently on shared hardware platforms has promoted building large scale computing infrastructures. As a result, data centers have been touted as one of the key enabling technologies for the fast growing IT industry and at the same time, resulting in a global market size of 17 billion US dollars in the first quarter of 2018. This represents a 47% increase on the same period in 2017 [2]. Data centers being large scale computing infrastructures have huge energy budgets, which have given rise to various energy efficiency issues. The energy consumed by a data center can be broadly categorized into two parts [3]: energy use by IT equipment (e.g., servers, networks, storage, etc.) and usage by infrastructure facilities (e.g., cooling and power conditioning systems). An approach to manage data center energy consumption consists of five main steps. The steps are parameter extraction, model construction, model validation, model reliability and application of the model to a task such as prediction.

Parameter extraction: In order to reduce the energy consumption of a data center, we first need to measure the energy consumption of its components [4] and identify where most of the energy is spent. This is the task of the parameter extraction phase.

Model construction: Second, the selected input parameters are used to build an energy consumption model using analysis techniques such as regression, machine learning, deep learning, etc. One of the key problems we face in this step is that certain important system parameters such as the power consumption of a particular component in a data center cannot be measured directly. Classical analysis methods may not produce accurate

results in such situations, and particle swarm optimization work better. The outcome of this step is a power model.

Model validation: Next, the model needs to be validated for its fitness for its intended purposes.

Model reliability: The repeatability and consistence of the process is the measure that indicate the variance of the component energy consumption.

Model usage: Finally, the identified model can be used as the basis for predicting the component or system's energy consumption. Such predictions can then be used to improve the energy efficiency of the data center, for example by incorporating the model into techniques such as temperature or energy aware scheduling [5], resource virtualization [6], improving the algorithms used by the applications [7], switching to low-power states [8], power capping [9], or even completely shutting down unused servers [10], etc. to make data centers more energy efficient.

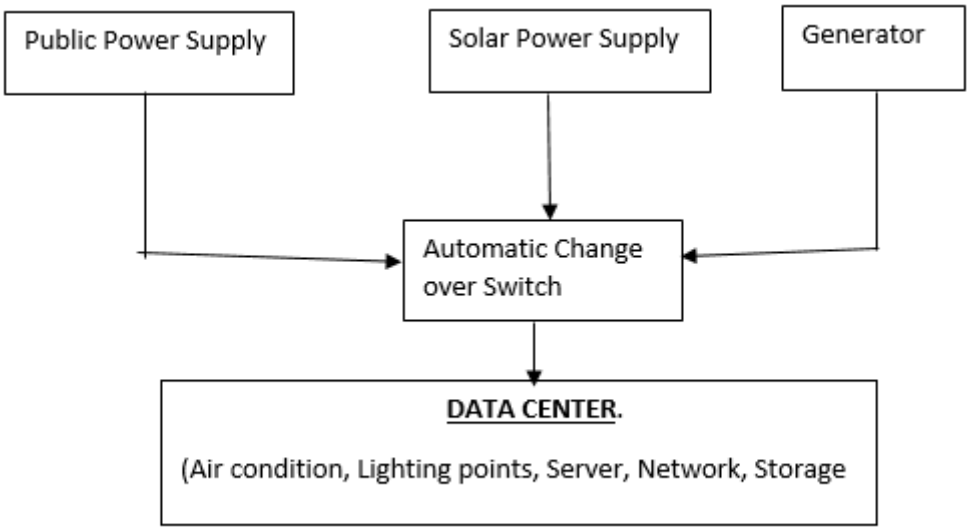

Figure 1: Energy flow block in a data center

2. Energy Consumption Models

The energy consumption considered in this paper includes the energy consumption of physical infrastructures and the energy consumption of IT equipment. Data centers are typically energized through the electrical grid. However, there are also data centers which use diesel, solar, wind power, etc. among other power sources. The electric power from external sources (i.e., the total facility power) is divided between the IT equipment, the infrastructure facilities. *Computer room air conditioning (CRAC)* units, a part of the cooling infrastructure, receive power through *uninterrupted power supplies (UPSs)* to maintain consistent cooling even during possible power failure. Note that certain power components such as battery backup may not be available in many data centers. The composition or layout of data center are as shown in fig 1.

A. IT Infrastructure Consumption

The energy consumption generated by the IT equipment primarily includes the energy consumption of the CPU, storage and network interface. Moreover, compared with other system resources, the CPU accounts for most of the energy consumption. Therefore, the energy consumption of the IT equipment need only consider the energy consumption of the CPU [12]. Studies have shown that the energy consumption of an IT equipment can be described by a linear function of its CPU utilization [13]. Additionally, studies have shown that the energy consumption of the idle IT equipment is approximately 70% of the fully CPU-utilized IT equipment [14], proving that shutting down an always idle IT equipment in a data center can reduce energy consumption. The energy consumption of the j_th server, storage and network system can be defined as (1).

$$E(ITMj) = \begin{cases} P_{idle} + (P_{busy} - P_{idle}) \times U_j^{cpu} & U_j^{cpu} > 0 \\ 0 & , U_j^{cpu} = 0 \end{cases} \quad (1)$$

where P_{idle} and P_{busy} represent the average energy consumption when the physical machine is idle and fully utilized. U_j^{cpu} represents the CPU utilization of the j_th physical machine.

One of the notable processor utilization based power models is the work by Fan *et al.* (appeared in the year 2007) which has influenced recent data center power consumption modeling research significantly. The research work has shown that the linear power model can track the dynamic power usage with a greater accuracy at the PDU level. If we assume the power consumed by a server is approximately zero when it is switched off, we can model the power *Pu* consumed by a server at any specific processor utilization as expressed in equation (1).

B. Non- IT Infrastructure Consumption

The non-IT infrastructure model is the consideration of data center power consumption through subsystems like cooling, lighting. Due to the activities of the millions and even billions of transistors integration in a serial or parallel chips known as processors and other related chips like north and south bridge chipsets, large volume of heat is generated in the center. Effective and efficient coordination of the working principles of the data centers, a cooling and illumination systems of optimal value are needed. The paper gears toward optimal use of the electrical energy without compromising quality of service. Maintaining adequate power quality levels and consistency of power supply is a must for effective data center operation. Non IT infrastructure system of a data center consumes significant amount of energy as the power wasted during transformation process which can be traced in its power hierarchy. Distribution of uninterrupted electrical power into a data center requires considerable infrastructure (such as transformers, changeover switch, cooling devices etc.)

In a typical data center power hierarchy, a primary switch board distributes power among multiple

Uninterrupted Power Supply sub-stations (UPSs). Each UPS in turn, supplies power to a collection of Power Distribution Units (PDUs). A PDU is associated with a collection of server racks and each rack has several chassis that host the individual servers. Such an arrangement forms a power supply hierarchy within a data center.

PDUs are responsible for providing consistent power supply for the servers. They transform the high voltage power distributed throughout the data center to voltage levels appropriate for servers. PDUs incur a constant power loss which is proportional to the square of the load which can be represented as in equation (2).

$$P_{pdu\_loss} = P_{pdu\_idle} + \pi_{pdu}(\sum_N P_{serv})^2 \quad (2)$$

where P_{pdu_loss} represents power consumed by the PDU, while π_{pdu} represents the PDU power loss coefficient, and P_{pdu_idle} which is the PDU's idle power consumption. The number of servers in the data center is represented by *N*.

Cooling systems are used to effectively maintain the temperature of a data center. Cooling power is the biggest consumer of the non-computing power in a data center followed by power conversion and other losses [16]. The data center cooling power is a function of many factors such as layout of the data center, the air flow rate, the spatial allocation of the computing power, the air flow rate, and the efficiency of the CRAC units.

Typically, the largest consumer of power and the most inefficient system in a data center is CRAC.

Factors that affect such operation of the CRAC unit include the operational efficiency and the air distribution design of the unit. Percentage of the cooling power varies, but can be up to 50% or more in a poorly designed and operated data center [17]. About 40% of the total energy consumption in the telecom industry is devoted to cooling equipment in data centers [18]. Most of this energy is consumed by the site chiller plant, CRAC, and by the air handlers. Heat dissipation in a data center is related to its server utilization. Studies have shown that for every 1W of power utilized during the operation of servers an additional 0.5–1 W of power is consumed by the cooling equipment to extract the heat out from the data center [19]. Power consumption of the data center cooling equipment generally depends on two parameters, first the amount of heat generated by the equipment

within the data center and second due to the environmental parameters such as temperature. Moore *et al.* [20] modeled the cooling power consumption as,

$$C = \frac{Q}{\eta(T=T_{sup}+T_{adj})} + P_{fan} \qquad (3)$$

where Q is the amount of server power consumption, $\eta(T = T_{sup} + T_{adj})$ is the η at $T_{sup} + T_{adj}$ Note that T_{sup} is the temperature of the cold air supplied by the CRAC units. They assumed a uniform T_{sup} from each CRAC unit. T_{adj} is the adjusted CRAC supply temperature. η is the coefficient of performance which gives the performance of the CRAC units.

3. Energy Consumption Optimization

As the heat dissipation from the servers in a data center ought to be proportional with cooling systems in the same center. But in most cases the disconnect or lack of communication between the entire network system and the cooling system contribute to large extent the energy losses or consumption recorded in the data center. Network servers works in two major modes: idle mode and operation mode. In this paper, the researcher investigates the losses in idle mode of the network servers of the data center. The configuration of servers was not discussed here but it was implied in this study. The research seeks to deploying the internet of things (IOT) for seamless connection, communications, monitoring and evaluation. In Fig.2 the IOT sensors are connected to monitor the server activities to be able to make an informed decision through structured feedback and control. The embedded system also helps in switching cooling system modules in response to the feedback.

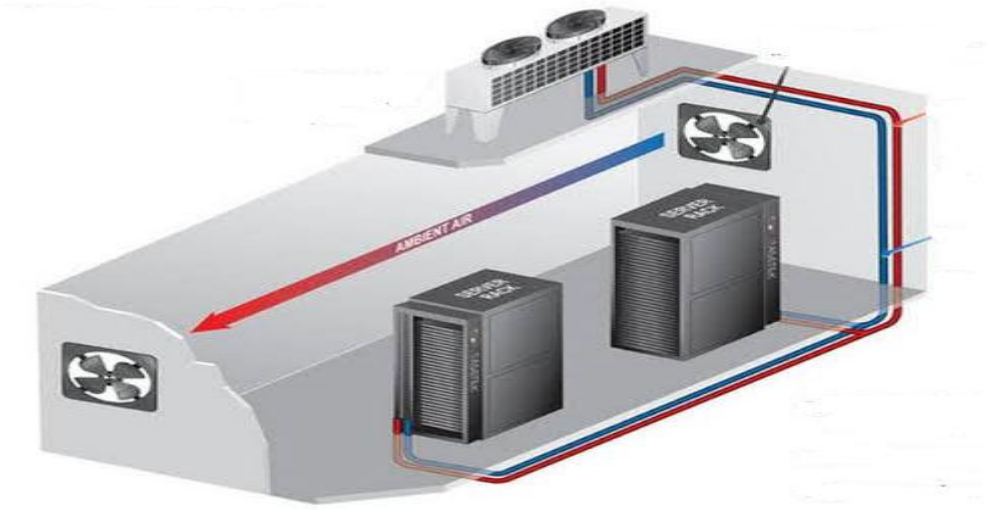

Figure 2: Data center setup with IOT devices

An evolutionary approach, Particle Swarm Optimization (PSO) was used to improve and optimize sensor networks of the cooling systems and the server real time activities reporting. The particle swarm optimization is a population based optimization technique, introduced by Kennedy and Eberhert in 1995. The model of this algorithm is based on the social behavior of bird flocking. It works through initializing population of random solutions and searching for the optima by updating generations. The PSO technique uses several particles, each represents a solution, and finds the best particle position with respect to a given fitness function. In this research, the PSO optimizes the sensors and server connectivity in the data centers.

In PSO, each single solution is a "bird" in the search space. We call it "particle". All of particles have fitness values that are evaluated by the fitness function to be optimized, and have velocities that direct the flying of the particles. The particles fly through the problem space by following the current optimum particles. PSO is initialized with a group of random particles (solutions) and then searches for optima by updating generations. In every generation, each particle is updated by following two "best" values. The first one is the best solution (fitness) it has achieved so far. (The fitness value is also stored.) This value is called 'pbest'.

Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the population. This best value is a global best and called gbest. After finding the two best values, the particle updates its velocity and positions. The particle swarm optimization concept consists of, at each time step, changing the velocity of (accelerating) each particle toward its pbest and gbest locations. Acceleration is weighted by a random term, with separate random numbers being generated for acceleration toward pbest and gbest locations.

In the past several years, PSO has been successfully applied in many research and application areas. It is demonstrated that PSO gets better results in a faster, cheaper way compared with other methods. Another reason that PSO is attractive is that there are few parameters to adjust. One version, with slight variations, works well in a wide variety of applications. Particle swarm optimization has been used for approaches that can be used across a wide range of applications, as well as for specific applications focused on a specific requirement.

Velocity and Position expression for PSO.

As the velocity of each particle dynamically changes according to $pBest$ and $gBest$ the velocity, V_{id}, is updated according to the following equation:

$$V_{id}^{new} = w \times V_{id}^{old} + c_1 \times R_1 \times (pBest_{id} - x_{id}) + c_2 \times R_2 \times (gBest_{id} - x_{id}) \quad (4)$$

After the velocity is updated, the position for a particle is also updated as stated below:

$$x_{id}^{new} = x_{id}^{old} + V_{id}^{old} \quad (5)$$

Where $pBest_{id}$ and $gBest_{id}$ represent Particle's best position Global best position respectively. W Inertia weight

x_{id} the current particle position and V_{id} the current particle velocity. $c1$, $c2$ are Learning factors and $R1$, $R2$ Random numbers between 0 and 1

4. Results

The results of the technique adopted for this research as simulated are expressed in this session. In Fig 3 the deployment of particle swarm optimization was visibly significant in energy savings. It indicates that as the number of active servers increases the corresponding cooling responses will also increase exponentially to some point. In the other case, or system without optimization mechanism the cooling system remain fairly constant irrespective of the active and idle mode operation of the server.

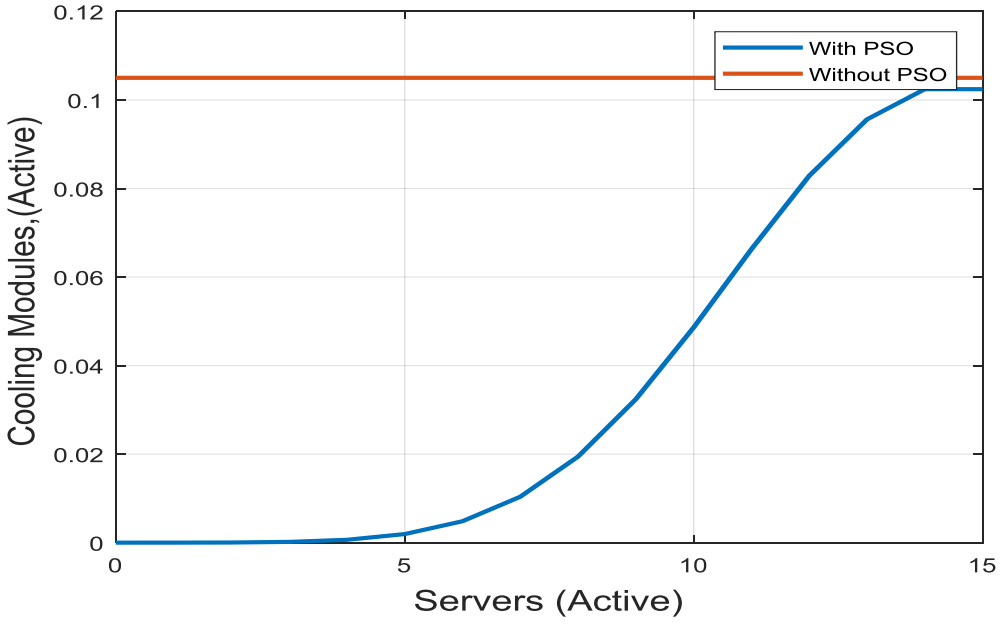

Figure 3: Server and cooling system operation of a data center

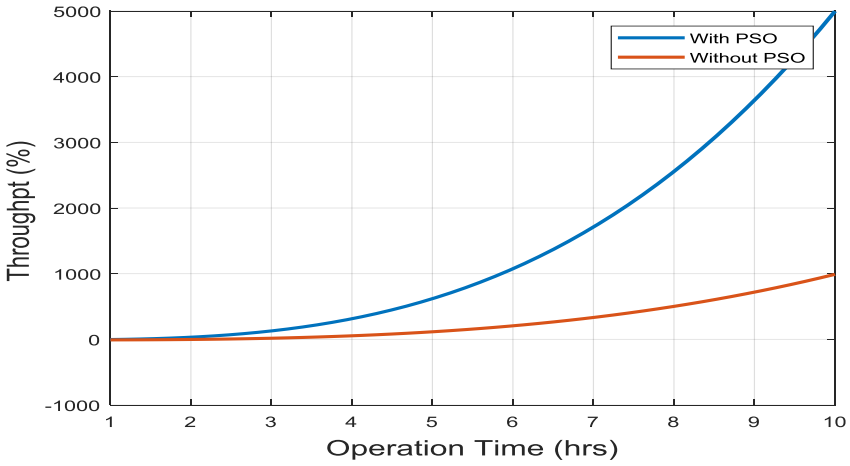

Figure 4: System throughput with regard to IOT operation

The activities of the server response were studied and it indicates that the throughput of the data center improved as shown in Fig 4. The index of the increase was purely depending on the optimization mechanism used on the research. In Fig 5 power consumption of the data center was also evaluated on daily basis. It shows that more power was consumed on the bit to maintain zero down time in the data center. But the same graph revealed the power savings made in average as a result of the integration of particle swarm optimization in the system. The quality of service were not compromised in any way, hence a better model for energy conservation.

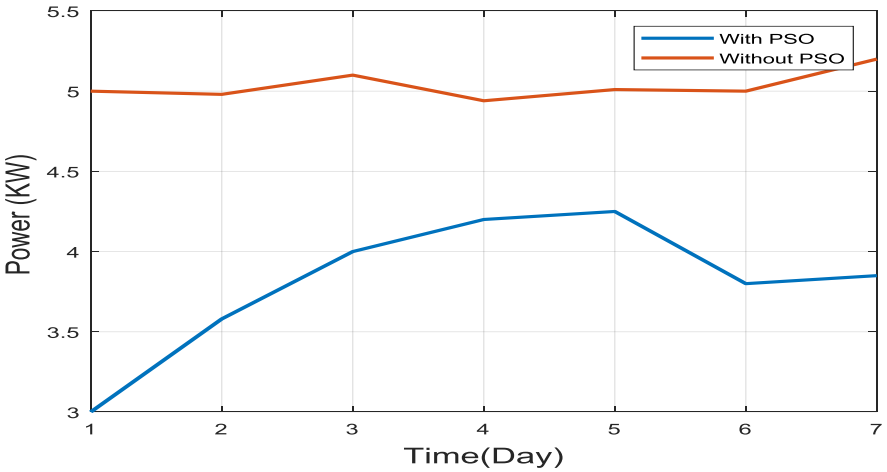

Figure 5: Power consumption of the data center

5. Conclusion

In this paper, particle swarm optimization technique has been effectively integrated to optimize energy consumption in a data center. The use of IOT made it possible through embedded subsystem for communication, feedback and control via high level automation. Much volume of energy was saved as idle state of the server was monitored and cooling system was configured to respond to the situation.

References

[1]. R. Buyya, A. Beloglazov, and J. H. Abawajy, "Energy-efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges," *CoRR*, vol. abs/1006.0308, 2010.

[2]. Canalys Palo Alto, Singapore Shanghai and Reading (UK), 2018. [Online]. Available:http://www.canalys.com/newsroom/cloud-infrastructure-market-grows-47-q1-2018-despite-underuse.

[3]. "Energy efficiency policy options for Australian and New Zealand data centres," The Equipment Energy Efficiency (E3) Program, 2014.
[4]. J. Brown and C. Reams, "Toward energy-efficient computing,"*Queue*, vol. 8, no. 2, pp. 30:30–30:43, Feb. 2010.
[5]. F. Bellosa, "The benefits of event: Driven energy accounting in power sensitive systems," in *Proc. 9th Workshop ACM SIGOPS EQ—Beyond PC—New Challenges Oper. Syst.*, 2000, pp. 37–42.
[6]. S.-Y. Jing, S. Ali, K. She, and Y. Zhong, "State-of-the-art research study for green cloud computing," *J. Supercomput.*, vol. 65, no. 1, pp. 445–468, Jul. 2013.
[7]. H. Liu, C.-Z. Xu, H. Jin, J. Gong, and X. Liao, "Performance and energy modeling for live migration of virtual machines," in *Proc. 20th Int. Symp. HPDC*, 2011, pp. 171–182.
[8]. Beloglazov and R. Buyya, "Energy efficient resource management in virtualized cloud data centers," in *Proc. 10th IEEE/ACM Int. CCGrid*, May 2010, pp. 826–831.
[9]. Feller, C. Rohr, D. Margery, and C. Morin, "Energy management in iaas clouds: A holistic approach," in *Proc. IEEE 5th Int. Conf. CLOUD Comput.*, Jun. 2012, pp. 204–212.
[10]. Lefurgy, X.Wang, and M.Ware, "Power capping: A prelude to power shifting," *Cluster Comput.*, vol. 11, no. 2, pp. 183–195, Jun. 2008.
[11]. M. Lin, A.Wierman, L. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," *IEEE/ACM Trans. Netw.*, vol. 21, no. 5, pp. 1378–1391, Oct. 2013.
[12]. Beloglazov, J. Abawajy, and R. Buyya, ``Energy-aware resource allocation heuristics for ef_cient management of data centers for cloud computing," *Future Generat. Comput. Syst.*, vol. 28, no. 5, pp. 755_768, 2012.
[13]. X. Fan, W. Weber, and L. Barroso, ``Power provisioning for a warehouse sized computer," in *Proc. 34th Annu. Int. Symp. Comput. Archit.*, 2007, pp. 13_23.
[14]. L. Hu, H. Jin, and X. Liao, ``Magnet: A novel scheduling policy for power reduction in cluster with virtual machines," in *Proc. IEEE Int. Conf. CLUSTER Comput.*, 2008, pp. 13_22.
[15]. X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *Proc. 34th Annu. ISCA*, 2007, pp. 13–23.
[16]. Z. Abbasi, G. Varsamopoulos, and S. K. S. Gupta, "TACOMA: Server and workload management in internet data centers considering cooling-computing power trade-off and energy proportionality," *ACM Trans. Archit. Code Optim.*, vol. 9, no. 2, pp. 11:1–11:37, Jun. 2012.
[17]. M. Patterson, "The effect of data center temperature on energy efficiency," in *Proc. 11th Intersoc. Conf. ITHERM Phenom. Electron. Syst.*,May 2008, pp. 1167–1174.
[18]. J. Dai, M. Ohadi, D. Das, and M. Pecht, "The telecom industry and data centers," in *Optimum Cooling of Data Centers*. New York, NY, USA: Springer-Verlag, 2014, pp. 1–8.
[19]. Z. Wang, N. Tolia, and C. Bash, "Opportunities and challenges to unify workload, power, and cooling management in data centers," *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 3, pp. 41–46, Aug. 2010.
[20]. J. Moore, J. Chase, P. Ranganathan, and R. Sharma, "Making scheduling 'cool': Temperature-aware workload placement in data centers," in *Proc. Annu. Conf. USENIX ATEC*, 2005, p. 5.