Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

NEW TRENDS AND IDEAS

iThermoFog: IoT-Fog based Automatic Thermal Profile Creation for Cloud Data Centers using Artificial Intelligence Techniques

Shreshth Tuli*[1] | Sukhpal Singh Gill[2] | Giuliano Casale[3] | Nicholas R. Jennings[3]

[1]Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Delhi, India
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
[3]Department of Computing, Imperial College London, London, UK

Correspondence
*Shreshth Tuli, Department of Computer Science and Engineering, IIT - Delhi 110016, India.
Email: shreshthtuli@gmail.com

Abstract

Preventing failures in Cloud Data Centers (CDCs) due to high temperatures is a key challenge. Such centers have so many servers that it is very difficult to efficiently keep their temperature under control. To help address this issue, we propose an artificial intelligence (AI) based automatic scheduling method that creates a thermal profile of CDC nodes using an integrated Internet of Things (IoT) and Fog computing environment called ***iThermoFog***. We use a Gaussian Mixture Model to approximate the thermal characteristics of the servers which are used to predict and schedule tasks to minimize the average CDC temperature. Through empirical evaluation on an iFogSim and ThermoSim based testbed and IoT based smart home application, we show that iThermoFog outperforms the current state-of-the-art thermal-aware scheduling method. Specifically, iThermoFog reduces mean square temperatures by 13.5%, while simultaneously improving energy consumption, execution time, scheduling time and bandwidth usage.

KEYWORDS:
Thermal-Aware; Artificial Intelligence; Fog Computing; ThermoSim; iFogSim; IoT; Cloud Data Centers

1 | INTRODUCTION

Prominent Cloud providers such as Facebook, Google, Microsoft and Amazon are utilizing Cloud Data Centers (CDC) to provide high quality services to Cloud users [1]. CDCs with large numbers of servers provide reliable Cloud services and fulfil the demands of users [2]. However, the heavy computational utilization of CDCs increases energy consumption and produces significant heat, which needs efficient cooling to keep temperatures under control [3]. Currently, temperature management techniques use reactive mechanisms to generate temperature profiles and control the cooling facilities available within the CDC [4]. But, as shown in prior work [5], such mechanisms can have large delays in profiling the thermal characteristics of the CDC, which rapidly increase the energy consumption and operational costs [6]. To solve this problem, a novel approach is required that is quickly able to adapt to dynamic task characteristics and host resource utilization in a heterogeneous environment that can not only keep temperature under control but also reduce energy, latency and bandwidth usage. To this end, we propose an Artificial Intelligence (AI) based automatic scheduling technique that uses the thermal-profile characteristics [5,7], to control the cooling facilities and maintain the CDC temperatures [8] using an integrated Internet of Things (IoT) [1] and Fog computing [9] environment. Specifically, our method creates the thermal-aware profile of the CDC using a Gaussian Mixture Model applied to past thermal measurements to predict and schedule tasks in order to minimize CDC temperature. Fog computing techniques are then used to control the temperature of the data center using edge devices and IoT sensors. To validate our technique, we implement a Cloud-Fog based environment

[0]**Abbreviations:** AI, Artificial Intelligence; IoT; Internet of Things; GTARA, Game based Thermal-Aware Resource Allocation

TABLE 1 Comparison of iThermoFog with Related Works

Work	Cloud	Fog	AI	IoT	Thermal Profile Creation	Implementation			QoS Parameters					
						CloudSim	iFogSim	ThermoSim	Energy	Latency	Bandwidth	Execution Time	Temperature	Scheduling Time
Liu et al.[12]	✓					✓			✓				✓	
Akbar et al.[13]	✓					✓			✓				✓	
Khaleel[14]	✓					✓			✓			✓		
Ilager et al.[4]	✓					✓			✓			✓	✓	
iThermoFog	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

before moving to the real testbed to test its feasibility and efficacy in heterogeneous Fog environments. We compare service characteristics like energy, latency, bandwidth consumption, scheduling time, execution time and average temperature with GTARA, a state-of-the-art thermal aware scheduler proposed by Akbar et al. Experiments show that iThermoFog is able to use Cloud+Fog to reduce energy, bandwidth usage, execution time, latency and mean square temperature by 12%, 4%, 9.3%, 5.6% and 6.1% respectively.

1.1 | Motivation and Our Contributions

Our AI-based automatic thermal profile creation and scheduling model, ***iThermoFog***, is the first method that is able to predict temperatures for different scheduling decisions for optimal schedules which minimize the temperature of the CDC. Further, we evaluate various performance metrics of our model on an integrated IoT-Fog-Cloud environment. An IoT-based smart home application is used to generate the data by running various time-critical computational applications including heart patient analysis[3] and smart-home management application[10]. To validate iThermoFog, we create a simulation environment using iFogSim[11] and ThermoSim[5]. The former is an open-source Fog simulator for running Fog applications on simulated Fog nodes and measuring task characteristics like latency, network bandwidth, energy, scheduling time, and execution time. The latter is a thermal-aware simulator that allows developers to obtain and utilize thermal characteristics of host machines in an iFogSim setup.

Section 2 discusses related work and describes how iThermoFog addresses the issue of temperature prediction and thermal-aware scheduling in time-critical applications. Section 3 describes the model and architecture of our thermal-profiling and scheduling framework. Section 4 compares the performance of various Fog-Cloud scenarios in terms of metrics such as execution time, network bandwidth consumption, energy consumption, latency and temperature. Finally, Section 5 concludes this work.

2 | RELATED WORK

Liu et al.[12] propose a thermal and power-aware model which jointly considers energy consumption arising from computing tasks, cooling, and task migrations. However, due to modeling limitations this work does not consider energy consumption arising from I/O processing and network transmission. Our model is more suitable for static workload cases and can be extended to dynamically update the scheduling policy for stochastic workloads. Akbar et al.[13] propose a game-theoretic thermal-aware allocation strategy (GTARA). Their work presents a methodology to efficiently manage the computational diversity within a Cloud data center by using the concept of cooperative game theory with Nash-bargaining to assign resources based on a thermal profile. A limitation of the approach is that the system is assumed to consist of homogeneous servers and does not consider automatic model update. Khaleel[14] describes a thermal-aware load balancing strategy that involves calculating the shortest distance to Cloud resources deployed at different geographical locations and conserving bandwidth, cost and energy using thermal characteristics. Further, the paper suggests that uniformly distributing workloads to different servers prevents computational hot-spots and maintains server health. This work however does not consider other types of utilization and energy consumption profiles arising from disks, network, and memory, which are critical to consider in Fog computing. Ilager et al.[4] propose an energy and thermal-aware scheduling algorithm (ETAS) that dynamically consolidates VMs to minimize the overall energy consumption while proactively preventing thermal hot spots. Another limitation is that the algorithm assumes a static cooling environment, which may not be versatile to different cooling settings. Table 1 compares iThermoFog with existing work based on the key parameters relevant in this work.

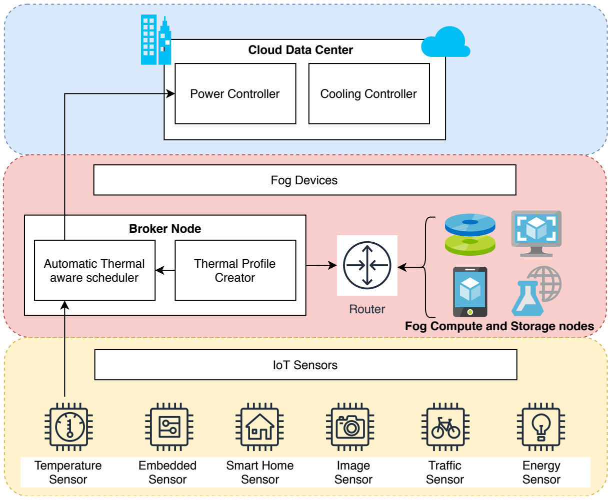

FIGURE 1 iThermoFog system model

3 | iThermoFog MODEL

Figure 1 shows the system architecture of iThermoFog, which consists of three different layers: Cloud, Fog, and IoT. The IoT layer consists of various pre-configured devices to gather data from other devices, such as sensors and smartphones. The thermal sensor measures the temperature of CDC servers periodically, which is helpful to generate the thermal profile automatically. The Fog layer consists of various Fog devices, such as the broker, and Fog nodes for fast computation. The broker acquires data from temperature sensors and creates the thermal profile. Subsequently, it schedules the Cloud resources using the thermal-aware scheduling technique introduced in ThermoSim[5]. Lightweight jobs are processed at Fog devices using available Fog storage and compute nodes to reduce latency, while heavy jobs are transferred to the Cloud layer for execution. The Cloud layer consists of two components: the cooling controller and the power controller. The former is an actuator that provides the required cooling to the CDC as per the created thermal profile. The latter reduces the overall temperature of the Cloud system. The overall temperature is modelled as the root mean square of the individual host node temperatures as done by Gill et al[5]. Based on empirical evaluations, RMS provides least net energy consumption in CloudSim. Other alternate modeling strategies include minimization of peak average temperature and cooling cost[15].

3.1 | Thermal Profile Creator and Scheduler

To create the thermal profile of the CDC, we use the popular AI strategy of formulating temperature as a Gaussian Mixture Model (GMM) of different task and Cloud host parameters[16]. As shown previously by Khosravi et al.[17], temperature characteristics of CDCs can be modeled accurately by GMMs. The task parameters include CPU, RAM, disk and bandwidth requirements and host parameters include CPU, RAM, disk and bandwidth availability. We model the temperature as a Gaussian variable of the task and host parameters. Thus, $\mathcal{T} \sim \mathcal{N}(x; \mu, \Sigma)$ where $\mathcal{T}_i$ is the temperature of host i ($i \in \{0, 1, 2, \ldots, n\}$) with n hosts in the CDC. Also, μ and Σ are parameters of the model with x as the task and host parameters like CPU, RAM, disk and bandwidth requirement/consumption including the task to host mapping.

Using genetic-based expectation maximization algorithms[18], we find the optimal parameter values μ^* and Σ^* for each host. For feasible solutions, we consider a truncated GMM, ignoring the negative samples. To generate the training data, we run various pre-existing scheduling policies like random allocation, ETAS[4], and others[14,12] to form a dataset of temperature values for different scheduling task and host parameters, and schedules. Using this dataset, we find μ^* and Σ^*, which we then deploy to predict the temperature of the all possible scheduling decisions (provided by CloudSim) at the current state, given by the iFogSim simulator, and take the one with least $\|\mathcal{T}\|$. This decision is taken whenever there is a new task in the system. For dynamic or stochastic workloads, this scheduling decision would have to be taken periodically, say every few minutes. When allocation of a running task changes, it would have to be migrated to the newly allocated host, investigation of implementation details is planned as part of future work.

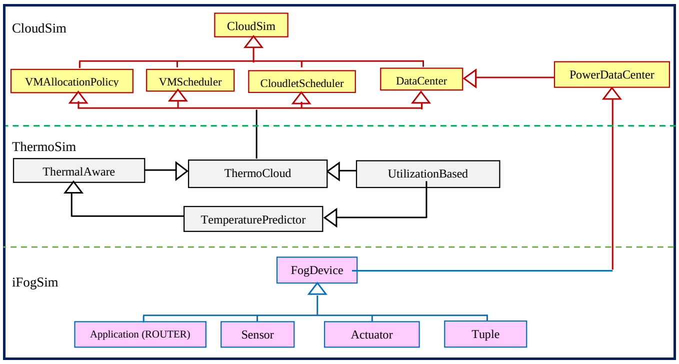

FIGURE 2 Integration of CloudSim, ThermoSim and iFogSim

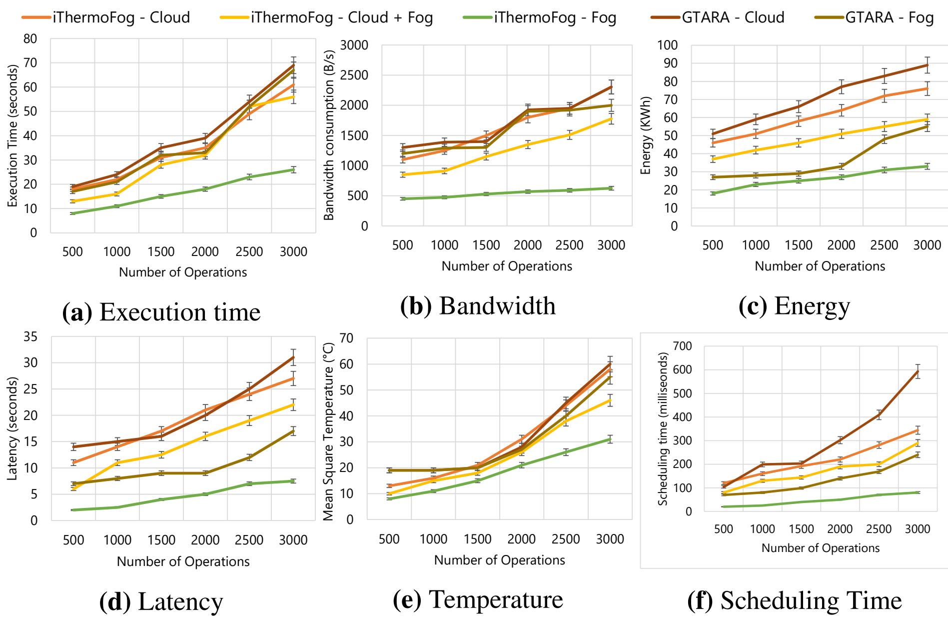

FIGURE 3 Comparison of iThermoFog with GTARA for different Cloud-Fog configurations

4 | PERFORMANCE EVALUATION

Experimental Setup: We have used iFogSim [11] and ThermoSim [5] to simulate the Fog environment, both of which work on top of CloudSim [19]. Integration of CloudSim, ThermoSim and iFogsim is described in Figure 2 . We use 12 cloud VMs and 18 fog devices in our setup, same as done in ROUTER application by Gill et al. [10]. As ThermoSim and CloudSim have been widely validated in previous work [4,3,10,11], we expect that these results could be reproduced in realistic settings.

DataSet: We use an IoT-based Smart Home Application called ROUTER [10], which describes the sequence of operations of an application and their type of tuples. The application modules are modeled in iFogSim using the AppModule class. There are data dependencies between modules, and these dependences are modeled using the AppEdge class in iFogSim. The control loop of interest for Smart Home application is also modeled in iFogSim using the AppLoop class. The application receives signals from different sensors and an actuator displays the current status of Smart Home to the user through Edge/Fog devices. The application model of the IoT-based Smart Home automation is built into iFogSim in order to validate the proposed technique through a real-time application. This means that the data from the experiment is directly fed into the simulator to provide edge-device operational behavior for the resource manager. We use the Particle Swarm Optimization (PSO) algorithm to schedule different operations on fog or cloud nodes as described by Gill et al [10]. We randomly create tasks of different ROUTER operation types and schedule on fog only nodes, cloud only nodes or both fog and cloud nodes based on the PSO approach.

Experimental Results: Figure 3 compares the performance of various QoS parameters such as energy, execution time, temperature, latency and network bandwidth for three different kind of services, i.e. Cloud, Cloud+Fog and Fog using iThermoFog. Figure 3 also shows that iThermoFog is able to reduce energy and bandwidth consumption, execution time, latency, and mean-squared temperature compared to GTARA[13]. As GTARA only works in the cloud layer, we also adapt it to work in Fog environments by considering all fog nodes same as cloud nodes, but with latency and computational characteristics as those of fog devices. Figure 3 (a) shows the variation of execution time with different numbers of operations. Upon increasing the number of operations, the execution time increases. The average value of the execution time in Fog is 14.44% and 18.92% less than less than Cloud+Fog and Cloud respectively. On an average, the execution time of iThermoFog is 52.68% and 9.35% faster than GTARA for Fog and Cloud respectively. The reason for the decrease in execution time is the automatic request handling mechanism of iThermoFog. Furthermore, iThermoFog tracks the state of all resources at each point of time, enabling it to take timely decisions. Figure 3 (b) further shows the average network bandwidth consumption for all the three different type of service. We see that Fog computing-based service consumes 12.42% less average network bandwidth than other two services and nearly 65% compared to the adapted GTARA. This is because iThermoFog processes lightweight data at edge devices effectively while fulfilling the deadlines dynamically. By increasing the number of operations, the energy consumption increases as shown in Figure 3 (c). The average value of energy consumption in the Fog computing environment is 11.32% and 15.72% less than Cloud+Fog and Cloud respectively. Compared to GTARA, iThermoFog reduces energy consumption by 32% for Fog and 12% for Cloud configurations. We see that the AI-based proactive scheduling of resources significantly reduces amount of network traffic, which leads to a reduction in the number of idle resources (processor, switching equipment, storage device, network device) that reduces the waste of energy. Further, we have analysed the latency of all the three services (i.e., the delay before transfer of user requests for job processing). With an increase in the number of operations, the value of latency increases as shown in Figure 3 (d). Here, latency is defined as the sum of scheduling time, execution time, communication and data transfer time from sensor to host. GTARA adapted to Fog setup is able to reduce latency significantly, however, iThermoFog in Fog configuration still outperforms adapted GTARA. It is observable that iThermoFog has a lower latency in contrast to the other two services. The average value of latency in iThermoFog is 12.76% and 16.91% less than Cloud+Fog and Cloud respectively. The reason is because iThermoFog executes job requests at Fog Data Server (FDS) instead of sending job requests to Cloud Data Server (CDS) which would result in a larger communication delay. Figure 3 (e) shows Fog offers 10.55% and 13.46% lower temperatures compared to Cloud+Fog and Cloud respectively. This is because it shuts down idle resources automatically. Finally, Figure 3 (f) shows how the scheduling time varies with number of operations. iThermoFog has low scheduling time compared to GTARA, as iThermoFog uses pre-computed GMM for temperature prediction compared to game-theoretic allocation. Low scheduling time (in milliseconds) leaves considerable margin to scale the technique in larger setups with low overheads.

5 | CONCLUSIONS AND FUTURE WORK

We have proposed an IoT and Fog computing environment based on an automatic thermal profile creation model which maintains the temperature of CDC proactively using GMM for thermal modeling. Further, the proposed model utilizes the past data and creates the thermal-aware profile using AI-based data analytics. An IoT based smart home application is used to generate data by performing different type of operations. Further, our iFogSim and ThermoSim based simulated Fog environment is used to validate the proposed model and optimize the QoS parameters such as latency, temperature, network bandwidth, energy and execution time.

The prominent future directions are described as follows: (1) The current model of iThermoFog can be extended to real deployments using frameworks like FogBus[3] and performance should be validated in unreliable and hybrid real systems with dynamic or stochastic workloads. (2) Analyze and enhance the scalability of the proposed model to allow large number of devices to be integrated without failures. (3) Implement dynamic/real-time offloading techniques for energy conservation in hybrid Fog-Cloud setups using iThermoFog. (4) Offloading using Mobile Cloud Computing for mobility and thermal-aware based offloading and task scheduling decision. (5) Investigate other datacenter temperature approximation methods.

References

1. Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems* 2009; 25(6): 599–616.

2. Gill SS, Tuli S, Xu M, et al. Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges. *Internet of Things* 2019; 8: 100118.

3. Tuli S, Mahmud R, Tuli S, Buyya R. FogBus: A Blockchain-based Lightweight Framework for Edge and Fog Computing. *Journal of Systems and Software* 2019; 154: 22–36.

4. Ilager S, Ramamohanarao K, Buyya R. ETAS: Energy and thermal-aware dynamic virtual machine consolidation in cloud data center with proactive hotspot mitigation. *Concurrency and Computation: Practice and Experience* 2019; 31(17): e5221.

5. Gill SS, Tuli S, Toosi AN, et al. ThermoSim: Deep Learning based Framework for Modeling and Simulation of Thermal-aware Resource Management for Cloud Computing Environments. *Journal of Systems and Software* 2020; 166: 1–20.

6. Sheikh HF, Ahmad I, Wang Z, Ranka S. An overview and classification of thermal-aware scheduling techniques for multi-core processing systems. *Sustainable Computing: Informatics and Systems* 2012; 2(3): 151–169.

7. Márquez AC, Del Castillo AC, Fernández JFG. Integrating artificial intelligent techniques and continuous time simulation modelling. Practical predictive analytics for energy efficiency and failure detection. *Computers in Industry* 2020; 115.

8. Wu X, Pellegrini FD, Gao G, Casale G. A Framework for Allocating Server Time to Spot and On-demand Services in Cloud Computing. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems* 2019; 4(4): 1–31.

9. Aljeri N, Boukerche A. Fog-Enabled Vehicular Networks: A New Challenge for Mobility Management. *Internet Technology Letters*: e141.

10. Gill SS, Garraghan P, Buyya R. ROUTER: Fog enabled cloud based intelligent resource management approach for smart home IoT devices. *Journal of Systems and Software* 2019; 154: 125–138.

11. Gupta H, Vahid Dastjerdi A, Ghosh SK, Buyya R. iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments. *Software: Practice and Experience* 2017; 47(9): 1275–1296.

12. Liu H, Liu B, Yang LT, et al. Thermal-aware and DVFS-enabled big data task scheduling for data centers. *IEEE Transactions on Big Data* 2017; 4(2): 177–190.

13. Akbar S, Malik SUR, Khan SU, Choo R, Anjum A, Ahmad N. A game-based thermal-aware resource allocation strategy for data centers. *IEEE Transactions on Cloud Computing* 2019.

14. Khaleel MI. Load Balancing and Thermal-Aware in Geo-Distributed Cloud Data Centers Based on Vlans. *Science Journal of University of Zakho* 2018; 6(3): 112–117.

15. Jiang X, Alghamdi MI, Al Assaf MM, et al. Thermal modeling and analysis of cloud data storage systems. *Journal of Communications* 2014; 9(4): 299–311.

16. Reynolds DA. Gaussian Mixture Models.. *Encyclopedia of biometrics* 2009; 741.

17. Khosravi A, Buyya R. Short-term prediction model to maximize renewable energy usage in cloud data centers. In: Springer. 2018 (pp. 203–218).

18. Pernkopf F, Bouchaffra D. Genetic-based EM algorithm for learning Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005; 27(8): 1344–1348.

19. Calheiros RN, Ranjan R, Beloglazov A, De Rose CA, Buyya R. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Software: Practice and experience* 2011; 41(1): 23–50.

How to cite this article: S. Tuli, S. S. Gill, G. Casale and N. Jennings (2020s), iThermoFog: IoT-Fog based Automatic Thermal Profile Creation for Cloud Data Centers using Artificial Intelligence Techniques , *Internet Technology Letters*, *2020;00:x–x*.