A Taxonomy and Capacity Planning Technique for Sustainable Cloud Computing – An Extensive Overview

Sivaraman Eswaran

Senior Lecturer, Dept. of Electrical and Computer Engineering, Curtin University, Miri, Malaysia

E-mail: sivaraman.eswaran@curtin.edu.my

Abstract

This overview of study intends to provide a thorough taxonomy of sustainable cloud computing capacity planning strategies. Several academic and industrial organizations have suggested several approaches to sustainability, and this taxonomy is used to analyze them. These modern methods have been analyzed and grouped together according to their shared traits and characteristics. This study takes a holistic look at sustainable Cloud Data Centers (CDCs), surveying the supporting methods and technologies along the way. It provides examples of successful capacity planning in sustainable CDCs based on research and practice from academia and industry. Moreover, the paper presents the most recent findings on what it takes to make CDCs viable. In addition, the difficulties of integration and the unanswered questions of sustainable CDC research have been discussed.

Keywords: Cloud computing, capacity planning, energy efficiency, cloud data centre, power management

1. Introduction

Innovations in Information and Communication Technology (ICT) have been exploded in the previous two decades. The ICT industry has become a major energy user and a factor in the rise in fuel costs, which has stymied the broad deployment of ICT equipment in some regions. The greatest expense in running such an establishment is the cost of electricity. Energy efficiency in the ICT industry has become more important as the sector's energy needs continue to grow, along with the associated high energy prices and depletion of natural resources. Over the last decade or more [1–6], researchers have paid a lot of attention to improving energy efficiency in this market. Although the estimates of net energy

DOI: https://doi.org/10.36548/jucct.2022.3.005
Received: 27.07.2022, received in revised form: 21.08.2022, accepted: 05.09.2022, published: 15.09.2022

consumption and per-device energy consumption may differ, all studies agree that the ICT sector's energy consumption will soon increase significantly.

Modular Data Centers (MDCs) may facilitate the energy as a location paradigm since they are transportable. In addition, by using virtual machine migration methods, the workload may be moved between strategically located, geographically scattered data centers in order to make use of renewable energy sources in other areas [7]. Figure 1 shows Taxonomy of MDC.

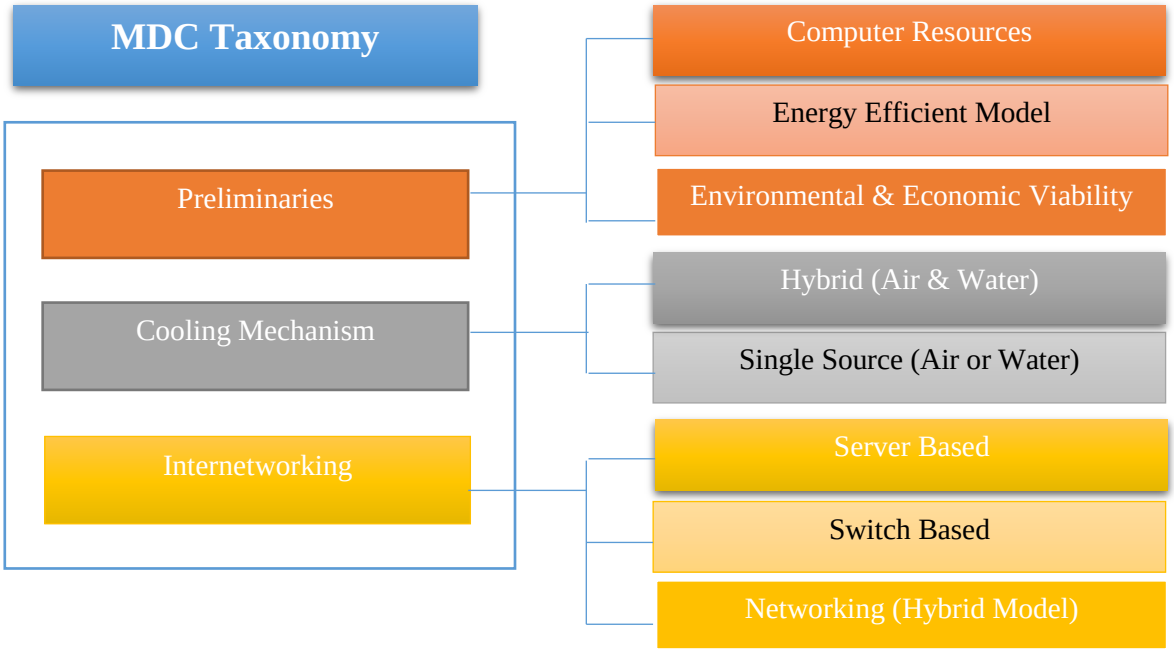

Figure 1. Taxonomy of Modular Data Centers for sustainability

The cloud computing model is a way of delivering and managing computer resources and services through the Internet on demand. In order to meet the ever-evolving needs of its customers, cloud computing must evolve into a more sustainable and energy-efficient model in the next years. Cloud service companies have difficulties in assuring the environmental and economic viability of their offerings. The sustainability of cloud services is also negatively impacted by the need for a large number of cloud datacenters, which drives up both prices and carbon footprints [8]. Figure 2 contains the data of early and recent years of capacity planning techniques.

When it comes to computing and it provides a scalable and robust computing environment through Internet-based subscription services. Many providers on cloud data centers provide the storage, processing, and connectivity needs of the Internet and other digital environments. There should be a backup set of computers in case one fails, so that the

service may continue without interruptions or delays in accordance with the Service Level Agreement.

Early Years	Recent Years
1. Trade – off between users 2. Analysis of Predictions for capacity planning 3. Generation of workload & its scheduling	1. Optimized Stochastic Model Generation 2. CDC service availability improvement 3. Generation of cost – optimal solution

Figure 2. Evolution of Capacity Planning Techniques

Although several studies have presented resource management strategies, algorithms, and architectures to address the issue of producing energy-efficient cloud services, this problem is still open for further study. New problems with resource scheduling may be overcome with the help of comprehensive resource management, which can guarantee a high degree of sustainability. Additionally, by improving waste heat usage and free cooling systems, cooling costs may be lowered [9-12]. Free cooling and renewable energy generation solutions can only work well if they are implemented in areas with favorable climatic conditions. To effectively execute waste heat recovery forecasts, it is also necessary to pinpoint where waste heat recovery may take place. CDCs may be moved depending on the below mentioned factors:

- The availability of green computing resources.
- The feasibility of waste heat recovery.
- Presence of free cooling resources.

2. Literature Survey

Stochastic Model-based Capacity Planning (SMCP) is a strategy presented by Ghosh et al., [13] for virtual infrastructure to meet the needs of users and carry out their workloads within predetermined time and cost constraints. Optical networks were installed to increase CDCs' carrying capacity and facilitate sustainability in CDCs, reducing energy use. For better capacity planning of cloud-based apps and associated QoS needs, Kouki et al. [14] developed the awareness technique. One of the greatest difficulties of the twenty-first century is the design of sustainable systems, which must take into account both the ecological transition and

the digital change. Additionally, determining the value of a strategy for enhancing a system's sustainability is difficult from a scientific standpoint. Sustainability encompasses four subfields of inquiry: ecology, sociology, technology, and economics.

To meet the needs of its customers, Carvalho et al. [15] developed a planning framework for capacity that would use several pricing models (including on-demand, reservation, and spot pricing) to meet their demands. CPF aids in determining the long-term price structure necessary to carry out present workloads. The primary focus of this meta-analysis is on the role of energy in the built environment and the economy. It addresses how to manage workloads and resources holistically in order to operate CDCs effectively with low energy (minimal energy cost). Sustainable cloud computing aims to do the following things:

- Lower energy use in data centers.
- Getting rid of old gadgets after they've served their purpose.

By relying on distant servers accessible over the Internet rather than on-premises hardware, cloud computing significantly boosts the pace of our economy. Users' data are stored, managed, and processed quickly and effectively thanks to the proliferation of data centers, but this comes at the cost of a larger carbon footprint and hence lower sustainability.

Capacity Planning for Cloud Infrastructure (CPCI) was the focus of Sousa et al.'s research [16], which sought to establish best practices by balancing the competing priorities of cost and reliability. The availability of servers was determined when creating a budget-friendly capacity plan with the use of a stochastic model generator. With proper resource management, cloud services may reduce their carbon footprint and improve their long-term viability.

Energy-efficient green data centers may be designed with the help of a technique described by Heller et al., [18] for Selecting Optimum Energy Sources (SOES) for sufficient capacity planning. The criterion for selecting energy sources is the capacity to design all major topics of investigation. In addition, SOES effectively lowers both operating and capital expenses throughout the course of a system's lifespan.

3. Sustainable Cloud Computing

High operating expenses and carbon emissions stem from the increased installation of several cloud datacenters. To successfully lower carbon emissions [19], CDCs in sustainable

cloud computing run on renewable energy sources, which replaces the traditional footprints that may be drastically reduced by the use of energy saving methods, thus contributing to cloud computing's sustainability [20]. Additionally, the CDCs are environmentally friendly since they make use of instruments for free cooling of the servers, and they recycle waste heat from the heat that the servers generate.

3.1 System Architectures

The cloud computing platform can manage the ever-increasing complexity of today's user applications. To cater to the vast variety of clients interested in eco-friendly computing, many application models are created for a variety of fields.

In a thread-based approach, numerous processes are broken down into smaller pieces called "threads," which run in parallel and share system resources like RAM, network bandwidth, and CPU. Large jobs are broken down into smaller tasks and run in parallel using dedicated cloud resources. Typical MapReduce jobs perform the mapped tasks by slicing the input dataset into discrete pieces that follow a consistent execution pattern. In addition, the maps' final results are sorted before being sent into the reduction functions as input [21].

3.2 Reduce Energy Consumption

Most of the time, increased performance and throughput take precedence over energy efficiency when designing new networks and data centers. There is a low-hanging fruit of opportunities to improve data center energy efficiency.

3.3 The Dynamic Management of Power Supply (DMPS)

The Elastic Tree, network-wide power management for data centers, was proposed early. The plan has just the connections that are currently being used by the system's workload, active at any particular time. The remaining connections are disabled. The data center workloads in their research exhibit predictable behavior during other times of the week, as shown by the traces collected from several data centers. For this reason, the method used past information to make predictions [17, 22].

Dynamic power management is a collection of methods for automatically turning off electronics in response to the changes in demand. Therefore, while still meeting the necessary performance and QoS standards, the number of active electrical components is kept to a minimum. When electronics are put into operation, they are usually subjected to a variety of

different workloads due to the interrelated nature of the devices. Future workload of these components may be estimated with some uncertainty using probabilistic approaches based on current workload history. Figure 3 shows the blocks of Renewable Energy based CDC.

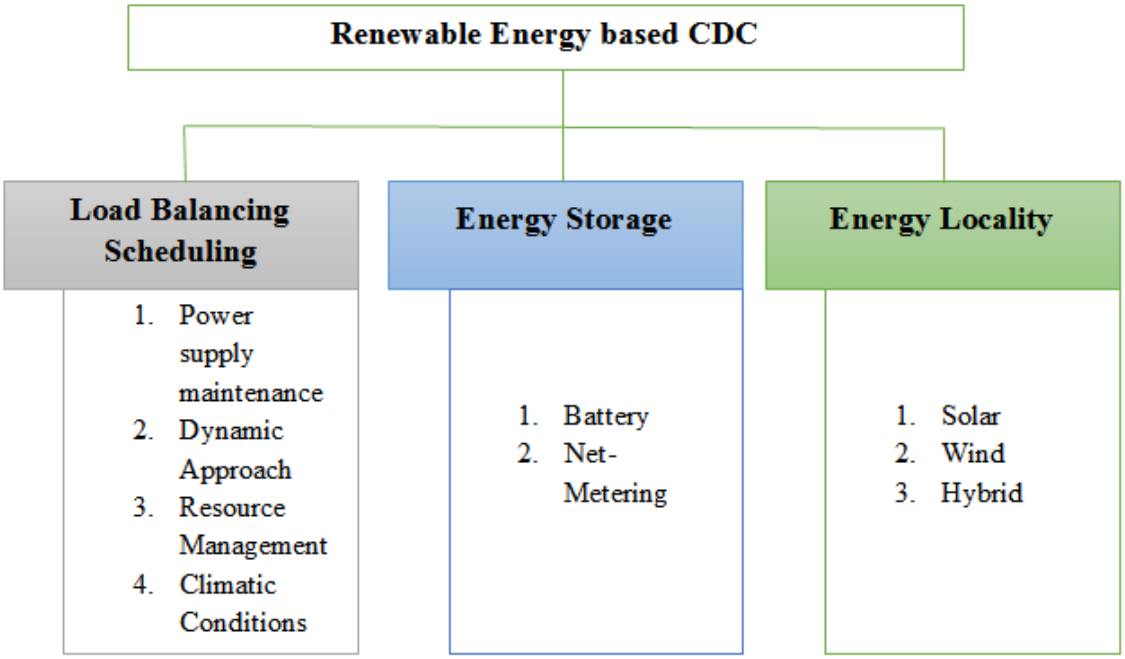

Figure 3. Renewable Energy based CDC

3.4 Capacity Planning

In order to facilitate environmentally responsible computing, planning for capacity may be done for the electricity grid, IT equipment, and air conditioning. A cloud datacenter's capabilities may be optimized by planning that takes end-user devices into account, particularly with regards to things like video-on-demand encoding. The Service Legal Agreement (SLA) should be in place for mission-critical characteristics like attract additional customers. Maximizing resource consumption through virtualization requires identifying which programs may be combined, which in turn necessitates considering critical utilization factors per application.

By maximizing resource usage and decreasing capacity costs, application consolidation contributes to the long-term viability of cloud computing systems [23], allowing for more effective capacity planning. Providing VM migration to transfer the least utilization of resources enhances the cloud data centers and is an essential part of any power infrastructure management strategy. Only with careful capacity planning can the cloud provide reliable service over time.

3.4.1 Methods for Managing Available Resources Evolution

To meet the needs of its customers, several pricing models (including on-demand, reservation, and spot pricing) are used. CPF aids in determining the long-term price structure necessary to carry out present workloads.

3.4.2 Capacity Planning Taxonomy

This function, and so on are all taken into account when making plans for capacity. Below, each of these taxonomy components are broken down and examples are provided when it makes sense to do so.

3.4.3 Subsystem

Every aspect of a cloud data center, from the IT equipment to the cooling and power systems, has to be carefully planned for capacity. Communication and Cloud Data Centers (CDCs) cannot function without the use of various IT equipment. In addition, an effective cooling strategy is necessary to keep the temperature of the cloud data center constant, which consumes a significant amount of energy. Last but not the least, a CDC can't function reliably if its power infrastructure hasn't been planned for.

3.4.4 Technology-Related Tasks

The capacity planning process takes into account both batch-based and crucial interactive IT workloads.

3.4.5 Possible Applications

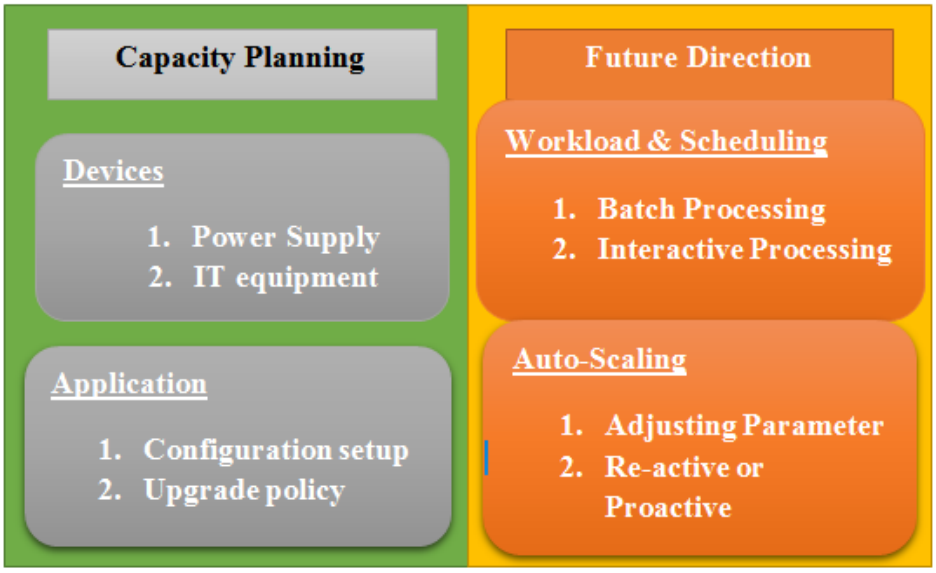

Figure 4. Capacity planning with Resources Evolution

Capacity planning that is both efficient and effective may be achieved using either the SLA-based or configuration-based design methods. The configuration-based model places more emphasis on violating SLAs. Figure 4 shows the capacity planning with resource evolutions in future.

3.4.6 Automatic Scaling

Autoscaling, either proactive or reactive, is planned for when a CDC is designed. To keep performance stable, reactive autoscaling uses feedback mechanisms to monitor and handle needs as they arise. In predicting and evaluating performance in terms of QoS values, proactive autoscaling stores the necessary capacity. Predictions have been recognized, and the appropriate steps to maximize CDC performance are being planned.

3.4.7 Functionality organizing throughout the system

In order to quantify, the many facets of capacity planning are established. The "cost" refers to the financial outlay necessary to develop a CDC of the specified make and model. A protocol's latency is the time it takes to execute under a certain CDC protocol configuration. In order to maximize the efficiency of cloud data centers, energy management of resources must be implemented prior to task execution. Capacity planning, which utilizes energy management of available resources, calls for management of power infrastructure [24-26].

4. Conclusion

Cloud computing is a relatively new paradigm that makes use of pre-existing technology to optimize the allocation of resources and the distribution of labour. This research work focuses on capacity planning in sustainable cloud computing for promoting environmental sustainability, and the key factors are as follows. The use of cloud computing should have negligible negative consequences on the environment, in accordance with the CDC's definition of sustainable practices. As part of their commitment to environmental protection, cloud service providers should reduce their use of non-renewable energy sources and replace them with renewable ones. Furthermore, this study provides important insights to the present studies on eco-friendly cloud computing that are split into numerous subfields, and are crucial for decision-makers, since it allows them to foresee the need for more thorough and realistic models for future computing technologies and mitigation strategies. In

addition, researchers may utilize the findings of this study to better understand the bounds and breadth of prior research on energy efficiency before developing new methodologies.

References

[1] Khosravi, Atefeh, and Rajkumar Buyya. 2018. Short-Term Prediction Model to Maximize Renewable Energy Usage in Cloud Data Centers. In Sustainable Cloud and Energy Services, pp. 203-218. Springer, Cham, 2018.

[2] Triantafyllidis, Charalampos P., Rembrandt HEM Koppelaar, Xiaonan Wang, Koen H. van Dam, and Nilay Shah. 2018. An integrated optimization platform for sustainable resource and infrastructure planning. Environmental Modelling & Software 101 (2018): 146-168.

[3] Rajkumar Buyya, and Sukhpal Singh Gill. 2018. Sustainable Cloud Computing: Foundations and Future Directions. Business Technology & Digital Transformation Strategies, Cutter Consortium, 21, 6, 1-10 (2018).

[4] Toni Mastelic, Ariel Oleksiak, Holger Claussen, Ivona Brandic, Jean-Marc Pierson, and Athanasios V. Vasilakos. 2015. Cloud computing: Survey on energy efficiency. ACM Computing Surveys, 47, 2 (2015): 1-33.

[5] Massimo Ficco, and Massimiliano Rak. 2016. Economic denial of sustainability mitigation in cloud computing. In Organizational Innovation and Change, Springer, Cham, 229-238, 2016.

[6] Xiang Li, Xiaohong Jiang, Peter Garraghan and Zhaohui Wu. 2018. Holistic energy and failure aware workload scheduling in Cloud datacenters. Future Generation Computer Systems, 78 (2018): 887-900.

[7] Fereydoun Farrahi Moghaddam, and Mohamed Cheriat. Sustainability-aware cloud computing using virtual carbon tax. 2015. arXiv preprint arXiv:1510.05182 (2015).

[8] Dejene Boru, Dzmitry Kliazovich, Fabrizio Granelli, Pascal Bouvry, and Albert Y. Zomaya. 2015. Energy-efficient data replication in cloud computing datacenters. Cluster computing 18, 1 (2015): 385-402.

[9] Muhammad Tayyab Chaudhry, Teck Chaw Ling, Atif Manzoor, Syed Asad Hussain, and Jongwon Kim. 2015. Thermal-aware scheduling in green datacenters. ACM Computing Surveys (CSUR) 47, 3 (2015): 1-39.

[10] Konstantinos Domdouzis. 2015. Sustainable cloud computing. Green Information Technology: A Sustainable Approach, Edited by: Mohammad Dastbaz, Colin Pattinson and Babak Akhgar (2015): 95-110.

[11] Josep Subirats, and Jordi Guitart. 2015. Assessing and forecasting energy efficiency on Cloud computing platforms. Future Generation Computer Systems 45 (2015): 70-94.

[12] Charith Perera, and Arkady Zaslavsky. 2014. Improve the sustainability of internet of things through trading-based value creation. In Proceedings of the World Forum on Internet of Things (WF-IoT), IEEE, 135-140. 2014.

[13] Rahul Ghosh, Francesco Longo, Ruofan Xia, Vijay K. Naik, and Kishor S. Trivedi. 2014. Stochastic model driven capacity planning for an infrastructure-as-aservice cloud. IEEE Transactions on Services Computing 7, 4 (2014): 667-680.

[14] Yousri Kouki, and Thomas Ledoux. "SLA-driven capacity planning for cloud applications. In Proceedings of the IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom), 2012: 135-140.

[15] Marcus Carvalho, Daniel A. Menascé, and Francisco Brasileiro. 2017. Capacity planning for IaaS cloud providers offering multiple service classes. Future Generation Computer Systems 77, (2017): 97-111.

[16] Erica Sousa, Fernando Lins, Eduardo Tavares, Paulo Cunha, and Paulo Maciel. 2015. A modeling approach for cloud infrastructure planning considering dependability and cost requirements. IEEE Transactions on Systems, Man, and Cybernetics: Systems 45, no. 4 (2015): 549-558.

[17] Fanxin Kong, and Xue Liu. "Greenplanning: 2016. Optimal energy source selection and capacity planning for green datacenters. In Proceedings of the ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS), 2016: 1-10.

[18] Heller B, Seetharaman S, Mahadevan P, Yiakoumis Y, Sharma P, Banerjee S, Mckeown N (2010) Elastictree: saving energy in data center networks. In: Proceedings of the 7th USENIX conference on networked systems design and implementation (NSDI'10). USENIX Association, Berkeley, CA

[19] Zhou Zhou, Zhi-gang Hu, Tie Song, and Jun-yang Yu. 2015. A novel virtual machine deployment algorithm with energy efficiency in cloud computing. Journal of Central South University 22, 3 (2015): 974-983.

[20] Sampaio, Altino M., and Jorge G. Barbosa. 2016. Energy-Efficient and SLA-Based Resource Management in Cloud Data Centers. Advances in Computers, Elsevier 100, 103-159, 2016.

[21] Charr, Jean-Claude, Raphael Couturier, Ahmed Fanfakh, and Arnaud Giersch. 2015. Energy consumption reduction with DVFS for message passing iterative applications on

heterogeneous architectures. In Proceedings of the IEEE International Parallel and Distributed Processing Symposium Workshop (IPDPSW), 922-931. IEEE, 2015.

[22] Claudio Fiandrino, Dzmitry Kliazovich, Pascal Bouvry, and Albert Zomaya. 2017. Performance and energy efficiency metrics for communication systems of cloud computing datacenters. IEEE Transactions on Cloud Computing 5, 4 (2017): 738-750.

[23] Sukhpal Singh Gill and Rajkumar Buyya, 2018. Failure Management for Reliable Cloud Computing: A Taxonomy, Model and Future Directions, IEEE Computing in Science and Engineering, 20, 4, 2018, 1-15.

[24] Li, Xiang, Peter Garraghan, Xiaohong Jiang, Zhaohui Wu, and Jie Xu. 2018. Holistic virtual machine scheduling in cloud datacenters towards minimizing total energy. IEEE Transactions on Parallel and Distributed Systems 29, 6 (2018): 1317-1331.

[25] Keke Gai, Meikang Qiu, Hui Zhao, and Xiaotong Sun. 2018. Resource Management in Sustainable Cyber-Physical Systems Using Heterogeneous Cloud Computing. IEEE Transactions on Sustainable Computing, 3, 2, 2018: 60-72.

[26] Tian Wang, Yang Li, Guojun Wang, Jiannong Cao, Md Zakirul Alam Bhuiyan, and Weijia Jia. 2017. Sustainable and Efficient Data Collection from WSNs to Cloud. IEEE Transactions on Sustainable Computing (2017). DOI: https://doi.org/10.1109/TSUSC.2017.2690301

Author's biography

Sivaraman Eswaran is currently working as a Senior Lecturer in the Dept. of Electrical and Computer Engineering, Curtin University, Miri, Malaysia. His area of research includes cloud computing, big data and networking.