Power System Technology
ISSN:1000-3673

Received: 06-10-2024 *Revised: 15-11-2024* *Accepted: 10-12-2024*

Optimizing Energy Consumption in Cloud Data Centers: A Survey of VM Allocation Techniques

Vipan*, Dr. Raj Kumar,
Department of Computer Applications, RIMT University, Mandi Gobindgarh, Punjab, India.
*Corresponding author, email: veekay88@gmail.com.

Abstract

The rapid expansion of cloud computing has resulted in a substantial increase in energy consumption, presenting a significant challenge for cloud service providers. To address this critical issue, researchers have proposed a wide array of innovative virtual machine allocation techniques aimed at optimizing energy usage in cloud data centers. This comprehensive survey examines these energy-efficient VM allocation approaches in detail, discussing their key methodologies, advantages, and limitations. By thoroughly investigating these techniques, this study aims to inform and guide future research and development efforts to drive more sustainable and efficient cloud computing infrastructures that can adapt to the growing demands of the cloud ecosystem. Specifically, this survey will delve deeper into the various energy-efficient VM allocation strategies, including dynamic VM consolidation, thermal-aware scheduling, and the integration of renewable energy sources. These techniques hold the potential to significantly reduce the energy demands of cloud computing and pave the way for more environmental friendly cloud infrastructures that can keep pace with the expanding needs of the cloud ecosystem.

Keywords: Cloud Computing, Data Centers, Energy Efficiency, Virtual Machine Allocation, Optimization, Green Computing.

1. Introduction

The rapid rise of cloud computing has positioned it as a transformative paradigm for delivering on-demand, scalable computing resources and services to users worldwide. This paradigm shift has fueled the proliferation of large-scale cloud data centers, which aim to cater to the growing global demand for cloud services. However, this expansion has come at a significant cost, resulting in a substantial increase in energy consumption and raising major concerns about the environmental impact and escalating operational expenses of these data centers[1][2]. International Energy Agency (IEA) , estimated that cloud computing data centers consumed around 2-3% of global electricity, i.e. 460 terawatt-hours (approximately) in 2022.The widespread adoption of cloud computing by individuals, businesses, and organizations has driven the rapid expansion of cloud data centers, leading to a corresponding surge in their energy consumption. These data centers now account for a substantial and ever-increasing portion of the world's total electricity supply, a concerning trend that is expected to continue as the reliance on cloud services grows [3][4]. Efficient management of cloud computing

Power System Technology
ISSN:1000-3673
Received: 06-10-2024 Revised: 15-11-2024 Accepted: 10-12-2024

resources, particularly the effective allocation and utilization of virtual machines, is crucial to addressing this pressing challenge.

Researchers have extensively studied and reported on energy-efficient VM allocation techniques, proposing a wide range of innovative approaches to optimize energy consumption in cloud data centers[5]. This paper aims to provide a comprehensive survey of these energy-aware VM allocation techniques, highlighting their key characteristics, advantages, and limitations, with the goal of informing future research and development efforts in this critical area. By delving deeper into the intricacies of these techniques, this survey will contribute to the ongoing efforts to address the escalating energy demands of cloud computing and pave the way for more sustainable and efficient cloud infrastructures.

Virtualization and VM Allocation

Virtualization is a fundamental enabling technology that underpins the cloud computing paradigm. It allows cloud service providers to consolidate and host multiple workloads on a single physical server through the deployment of virtual machines [6]. The process of allocating and mapping these virtual machines to physical servers, known as VM allocation , plays a crucial role in determining the overall efficiency and performance of cloud data centers. The efficient allocation of VMs to physical servers is essential for optimizing a range of key factors , including energy consumption, resource utilization, and application performance in cloud environments. Effective VM allocation strategies can help cloud providers maximize the benefits of virtualization by striking the right balance between resource consolidation, energy efficiency, and service-level requirements. This involves carefully considering factors such as the workload characteristics, resource utilization patterns, and the specific VM allocation algorithms employed [7].

Improper VM allocation can lead to suboptimal resource utilization, excessive energy consumption, and degraded application performance. Cloud providers must therefore invest in developing and deploying sophisticated VM allocation techniques that can dynamically adapt to changing workload conditions and optimize for various operational objectives, including energy efficiency , cost-effectiveness, and quality of service. By doing so, they can unlock the full potential of virtualization and deliver more sustainable and efficient cloud computing infrastructures[8].

Cloud Data Center Energy Usage

As cloud computing continues to grow, the energy consumption of cloud data centers is expected to rise significantly. Addressing this challenge is crucial, as cloud data centers now account for a substantial and ever -increasing portion of the world's total electricity supply [9] . Energy-efficient VM allocation techniques play a pivotal role in optimizing resource utilization and minimizing the energy footprint of cloud data centers. The key factors influencing energy consumption in cloud data centers include server utilization, cooling infrastructure, and network infrastructure [9]. To address these factors, researchers have explored various strategies, such as dynamic VM consolidation, thermal-aware scheduling, and the integration of renewable energy sources [10]. These techniques have the potential to significantly reduce the energy demands of cloud computing. Additionally , researchers have

Power System Technology
ISSN:1000-3673
Received: 06-10-2024 Revised: 15-11-2024 Accepted: 10-12-2024

developed a range of other energy-efficient VM allocation approaches, including load balancing, consolidation-based VM allocation, and dynamic VM allocation strategies [11]. Energy-aware, thermal-aware, and power-aware VM allocation algorithms have been proposed to optimize energy consumption, while QoS-oriented and workload-based VM allocation models aim to balance performance and efficiency . Renewable energy-driven VM allocation and heterogeneous hardware VM allocation approaches have also been investigated [11].

Furthermore, researchers have explored multi- objective VM allocation optimization, considering factors like energy, resource utilization, and performance [12]. Both distributed and centralized VM allocation architectures, as well as adaptive and predictive VM allocation mechanisms, have been developed and evaluated experimentally [12] . These advancements in VM allocation techniques are crucial for delivering more sustainable and efficient cloud computing infrastructures.

Challenges in VM Allocation

Effectively managing the allocation of virtual machines in cloud data centers poses several key challenges. Optimizing energy consumption is of paramount importance, as the growing reliance on cloud computing has led to a significant expansion in the energy footprints of data centers . Ensuring efficient utilization of critical resources, such as CPU, memory, and storage, is also crucial to maximizing the benefits of virtualization [13][14]. Maintaining the desired levels of application performance while balancing energy efficiency is a delicate tradeoff that requires careful consideration [6][3]. Furthermore, the dynamic and unpredictable nature of workloads in cloud environments presents a unique challenge. Cloud service providers must have the capability to effectively allocate and reallocate VM s to physical servers in response to changing resource demands, all while minimizing the impact on application performance and energy consumption [3][14]. Achieving this balance is essential for delivering reliable and cost-effective cloud services to end-users. Virtualization enables server consolidation, which can lead to significant energy savings. However, the potential overhead in energy usage and throughput reduction due to virtualization must be thoroughly evaluated and addressed[6]. Careful management of the virtualization layer and its impact on the underlying hardware is necessary to maximize the benefits of server consolidation while minimizing the drawbacks .

Impact of Virtualization on Energy Efficiency

Virtualization is a core component of cloud computing, enabling the consolidation of multiple virtual machines on a single physical machine. This consolidation can lead to significant energy savings by reducing the number of active physical servers required to meet the overall computational demand. However , the literature also suggests that the use of system- level efficiency techniques in clusters and grids might paradoxically increase their overall energy consumption in some cases[15]. This apparent contradiction highlights the complex trade-offs involved in optimizing energy efficiency in virtualized environments . While server consolidation through virtualization can yield substantial energy savings, the specific implementation and management of the virtualization layer can have a significant impact on the actual energy footprint of the overall system[16].Careful analysis and optimization of the virtualization process, along with a deep understanding of the underlying hardware and

Power System Technology
ISSN:1000-3673
Received: 06-10-2024 Revised: 15-11-2024 Accepted: 10-12-2024

workload characteristics, are crucial for maximizing the energy efficiency benefits of cloud computing.

The potential energy savings achieved through virtualization are influenced by various factors, including the workload characteristics , resource utilization patterns, and the specific VM allocation strategies employed.

2. Energy-Aware VM Allocation Techniques

Energy-aware VM allocation techniques can be further categorized as follows:

Static VM Allocation

These techniques make VM placement decisions based solely on the initial resource requirements of the VMs, without accounting for dynamic changes in workload over time. This approach is relatively simple to implement , as it does not require continuous monitoring and adjustment of VM placement. However, it may fail to adapt to fluctuations in resource demands, potentially leading to suboptimal resource utilization and energy efficiency [3]. Static VM allocation techniques are typically suitable for workloads with relatively stable resource requirements, but they may struggle to handle highly dynamic and unpredictable cloud workloads effectively.

Dynamic VM Allocation

These techniques continuously monitor the resource utilization of VMs and physical servers, and perform dynamic VM consolidation to optimize energy consumption [1].One example of a dynamic VM allocation technique is presented in [6]. The authors propose a mathematical model that accounts for the mutual influence among VMs to reduce resource contention, leading to an optimal scheduling of VMs under a given energy budget. In contrast to static allocation, dynamic VM allocation strategies address the need for reallocating VMs to new physical servers in response to changing workload conditions [3]. These dynamic approaches often involve live VM migration, where VMs are moved between physical servers without disrupting running applications.Dynamic VM allocation can be further divided into two main categories:

Proactive VM Allocation Techniques: Proactive VM allocation strategies are designed to anticipate future workload changes and proactively reallocate virtual machines to optimize for both energy efficiency and performance. These techniques leverage predictive models and workload forecasting to estimate future resource requirements, allowing the system to preemptively adjust the VM placement across physical servers. By proactively consolidating VMs onto a smaller number of active servers and selectively powering down underutilized hosts, proactive allocation approaches can achieve significant energy savings. At the same time, these strategies aim to maintain the desired application performance by ensuring that VMs are allocated to physical resources that can adequately support the predicted workloads. The proactive nature of these techniques enables cloud data centers to stay ahead of fluctuations in resource demands, optimizing the tradeoff between energy consumption and performance in a more holistic and anticipatory manner compared to reactive approaches[1].

Power System Technology
ISSN:1000-3673
Received: 06-10-2024 Revised: 15-11-2024 Accepted: 10-12-2024

***Reactive VM Allocation Techniques*:**These techniques dynamically respond to real-time changes in workloads, continuously monitoring the resource utilization of VMs and physical servers and performing live VM migrations as needed to maintain the desired performance and energy efficiency levels [14]. Unlike static allocation approaches, reactive techniques do not rely solely on the initial resource requirements of VMs, but rather adapt the VM placement across physical servers based on the fluctuating workload conditions. One example of a reactive VM allocation technique is presented by Jin et al., who implemented an approach that incorporates a speed switch and VM consolidation, considering both energy efficiency and response time [14]. This dynamic strategy aims to optimize the tradeoff between power consumption and application performance by adjusting the server speeds and consolidating VMs as the workload changes. Another reactive technique is the updated smart elastic scheduling algorithm developed by Kaur et al., which clusters VMs for migration to achieve a load-balanced allocation [14]. SESA continuously monitors the resource utilization of the VMs and the physical servers, and proactively migrates VMs to maintain an even distribution of the workload across the data center's infrastructure. This reactive approach helps to prevent hotspots and ensure efficient resource utilization, ultimately contributing to the overall energy efficiency of the cloud environment.

Renewable Energy-Aware VM Allocation: These techniques leverage the availability of renewable energy sources, such as solar and wind, to further optimize the energy consumption of cloud data centers. These approaches aim to align the data center's energy consumption with the supply of clean, renewable energy.These techniques often involve dynamic VM migration and server consolidation strategies to adapt the VM placement in real-time as the renewable energy supply fluctuates. By closely monitoring the renewable energy availability and dynamically adjusting the VM allocation, these approaches can minimize the reliance on grid-supplied electricity from fossil fuels.For instance, Mashayekhy et al. developed a VM allocation model that considers the availability and generation patterns of renewable energy sources, adjusting the VM placement accordingly. This approach aims to maximize the utilization of renewable energy while minimizing the overall energy costs and environmental impact of the data center's operations [17][18].

3. Load Balancing VM Allocation

Load balancing techniques aim to distribute the workload across multiple physical servers, preventing individual servers from becoming overloaded and ensuring efficient resource utilization. However, purely load-based VM allocation can lead to suboptimal energy consumption, potentially resulting in a larger number of active servers compared to consolidation-based approaches [1]. To address this, researchers have developed consolidated-based VM allocation strategies that aim to pack VMs onto the fewest number of physical servers possible, thereby minimizing the number of active servers and reducing overall energy consumption.[5]These consolidated-based approaches often incorporate load-balancing mechanisms to ensure that the workload is still distributed evenly across the active servers, preventing hotspots and maintaining application performance[19].

A wide range of VM allocation techniques have been explored in the literature, each with its own strengths, weaknesses, and target optimization objectives. Workload-based VM allocation

Power System Technology
ISSN:1000-3673
Received: 06-10-2024 Revised: 15-11-2024 Accepted: 10-12-2024

models analyze the resource demands of individual VMs and allocate them to physical servers accordingly, seeking to balance performance and efficiency[20].

4. Consolidation-based VM Allocation

Consolidation-based VM allocation techniques focus on packing VMs onto the fewest number of physical servers to reduce the overall energy consumption of the data center. These approaches often leverage mathematical optimization models, heuristic algorithms, or machine learning techniques to determine the optimal VM-to-server mapping[21].For example, Nasim et al. [3] proposed a mathematical model that considers the mutual influence among VMs to reduce resource contention and calculate the optimal VM scheduling under a given energy budget. Hwang et al. developed a hierarchical resource management solution that takes into account the correlations between VMs and the network topology to maximize the overall benefit [3].

5. Thermal-aware VM Allocation Approaches

Thermal-aware VM allocation techniques consider the thermal characteristics of the physical servers and aim to distribute VMs in a way that optimizes the cooling infrastructure's efficiency, reducing the overall energy consumption[22]. Thermal-aware VM allocation techniques consider the thermal characteristics of the physical servers and aim to distribute VMs in a way that optimizes the cooling infrastructure's efficiency, reducing the overall energy consumption[1]. For example, Gamal et al. developed an updated smart elastic scheduling algorithm that uses CPU utilization and RAM as parameters to cluster VMs on the same physical machine, considering load as a primary factor in the VM allocation process [14].

6. Power-aware VM Allocation Methods

Power-aware VM allocation methods focus on reducing the overall power consumption of the cloud data center by considering the power characteristics of the physical servers and the VMs.

These approaches often involve techniques such as server consolidation, dynamic voltage and frequency scaling, and the strategic placement of VMs to minimize the number of active servers. Serrano et al. proposed a power-aware VM allocation strategy that leverages a hybrid approach, combining consolidation-based and load-balancing techniques to optimize both energy efficiency and application performance[23].

7. QoS-oriented VM Allocation Techniques

QoS-oriented VM allocation techniques aim to ensure that the quality of service requirements of the hosted applications are met, while also optimizing for energy efficiency and other metrics. These approaches often involve mechanisms for monitoring and predicting application performance, as well as techniques for dynamically adjusting the VM allocation in response to changing workload conditions. For example, Feller et al. developed a VM allocation framework that considers both energy efficiency and application performance, using a combination of load forecasting and consolidation-based techniques to optimally place VMs[1].

Power System Technology
ISSN:1000-3673
Received: 06-10-2024 Revised: 15-11-2024 Accepted: 10-12-2024

8. Heterogeneous Hardware VM Allocation

Heterogeneous hardware VM allocation strategies aim to optimize the placement of VMs on diverse physical hardware configurations, such as servers with varying CPU, memory, and storage capabilities. These approaches seek to leverage the unique characteristics of different hardware components to improve energy efficiency, application performance, and resource utilization [24]. One such example is the work by Hwang and Pedram, who proposed a hierarchical resource management solution that considers the correlations between VMs and the network topology among the physical servers to maximize the overall benefit while minimizing energy consumption [25].

9. Green Computing VM Allocation Policies

Green computing VM allocation policies focus on reducing the environmental impact of cloud data centers by optimizing energy consumption and leveraging renewable energy sources. These policies often involve a combination of techniques, such as server consolidation, VM migration, and the strategic placement of VMs on physical servers that are powered by renewable energy sources. By strategically allocating VMs to physical servers in a way that aligns with the availability of renewable energy, these policies can help to reduce the reliance on grid-supplied electricity generated from fossil fuels and minimize the carbon footprint of cloud computing operations [26]. For example, Gao et al. developed a renewable energy-aware VM allocation model that considers both the availability of renewable energy and the energy efficiency of the physical servers to optimize the placement of VMs in a sustainable manner [17] [27]. The research in this field has shown that there are various approaches to optimizing energy consumption in cloud data centers, each with its own strengths and trade-offs.

10. Conclusion

This survey has offered a comprehensive overview of the diverse VM allocation techniques that have been developed to optimize energy efficiency in cloud data centers. The approaches discussed span a broad spectrum, encompassing mathematical models, heuristic algorithms, machine learning methods, thermal-aware allocation, power-aware allocation, QoS-oriented allocation, workload-based allocation, heterogeneous hardware allocation, and green computing policies. These techniques aim to reduce energy consumption, enhance resource utilization, and ensure the quality of service for hosted applications in cloud environments. As cloud computing continues to expand, the need for energy-efficient data center management will become increasingly crucial. This paper highlights the significant advancements made in this field and the promising potential for further innovations in optimizing energy consumption in cloud data centers through novel VM allocation strategies.

References

[1] A. Beloglazov, J. Abawajy, and R. Buyya, “Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing,” May 01, 2012, Elsevier BV. doi: 10.1016/j.future.2011.04.017.

Power System Technology
ISSN:1000-3673
Received: 06-10-2024 Revised: 15-11-2024 Accepted: 10-12-2024

[2] F. P. Tso, S. Jouët, and D. P. Pezaros, "Network and server resource management strategies for data centre infrastructures: A survey," Jul. 05, 2016, Elsevier BV. doi: 10.1016/j.comnet.2016.07.002.
[3] Y. Huang, H. Xu, H. Gao, X. Ma, and W. Hussain, "SSUR: An Approach to Optimizing Virtual Machine Allocation Strategy Based on User Requirements for Cloud Data Center," Jun. 01, 2021, Institute of Electrical and Electronics Engineers. doi: 10.1109/tgcn.2021.3067374.
[4] M. J. Usman et al., "Energy-efficient Virtual Machine Allocation Technique Using Flower Pollination Algorithm in Cloud Datacenter: A Panacea to Green Computing," Mar. 01, 2019, Elsevier BV. doi: 10.1007/s42235-019-0030-7.
[5] W. Sun, Y. Wang, and S. Li, "An optimal resource allocation scheme for virtual machine placement of deploying enterprise applications into the cloud," Jan. 01, 2020, American Institute of Mathematical Sciences. doi: 10.3934/math.2020256.
[6] Y. Jin, Y. Wen, Q. Chen, and Z. Zhu, "An Empirical Investigation of the Impact of Server Virtualization on Energy Efficiency for Green Data Center," Feb. 20, 2013, Oxford University Press. doi: 10.1093/comjnl/bxt017.
[7] W. Attaoui and E. Sabir, "Multi-Criteria Virtual Machine Placement in Cloud Computing Environments: A literature Review," arXiv (Cornell University). Cornell University, Jan. 01, 2018. doi: 10.48550/arxiv.1802.05113.
[8] T. M. Мирзоев and R. M. Alvarez, "Leveraging VMware vCloud Director Virtual Applications (vApps) for Operational Expense (OpEx) Efficiency," Jan. 01, 2014, Cornell University. doi: 10.48550/arxiv.1404.2157.
[9] M. Bansal, S. K. Malik, S. K. Dhurandher, and I. Woungang, "Policies and mechanisms for enhancing the resource management in cloud computing: a performance perspective," Jan. 01, 2020, Inderscience Publishers. doi: 10.1504/ijguc.2020.107615.
[10] N. Akhter,M. Othman,R. Naha, "Evaluation of Energy-efficient VM Consolidation for Cloud Based Data Center - Revisited." Oct. 2023. Available: https://arxiv.org/pdf/1812.06255.pdf
[11] Qiheng Zhou,Minxian Xu,S. S. Gill,Chengxi Gao,Wenhong Tian,Chengzhong Xu,R. Buyya, "Energy Efficient Algorithms based on VM Consolidation for Cloud Computing: Comparisons and Evaluations." Oct. 2023.
[12] C. Stier, J. Domaschka, A. Koziolek, S. Krach, J. Krzywda, and R. Reussner, "Rapid Testing of IaaS Resource Management Algorithms via Cloud Middleware Simulation," Mar. 30, 2018. doi: 10.1145/3184407.3184428.
[13] T. Khan, W. Tian, and R. Buyya, "Machine Learning (ML)-Centric Resource Management in Cloud Computing: A Review and Future Directions," arXiv (Cornell University). Cornell University, Jan. 01, 2021. doi: 10.48550/arxiv.2105.05079.
[14] A. Kaur et al., "Algorithmic Approach to Virtual Machine Migration in Cloud Computing with Updated SESA Algorithm," Jul. 03, 2023, Multidisciplinary Digital Publishing Institute. doi: 10.3390/s23136117.
[15] M. Zakarya and L. Gillam, "Energy efficient computing, clusters, grids and clouds: A taxonomy and survey," Mar. 16, 2017, Elsevier BV. doi: 10.1016/j.suscom.2017.03.002.

Power System Technology
ISSN:1000-3673
Received: 06-10-2024 Revised: 15-11-2024 Accepted: 10-12-2024

[16] S. Dharshika and N. G. Cholli, "Green Cloud Computing: Redefining the future of Cloud Computing," Jul. 01, 2021. doi: 10.47392/irjash.2021.203.
[17] R. Buyya and S. S. Gill, "Sustainable Cloud Computing: Foundations and Future Directions," Jan. 01, 2018, Cornell University. doi: 10.48550/arxiv.1805.01765.
[18] X. Wang, G. Zhang, M. Yang, and L. Zhang, "Green-Aware Virtual Machine Migration Strategy in Sustainable Cloud Computing Environments," in InTech eBooks, 2017. doi: 10.5772/67350.
[19] A. Paya and D. C. Marinescu, "Energy-aware Application Scaling on a Cloud," Jan. 01, 2013, Cornell University. doi: 10.48550/arxiv.1307.3306.
[20] R. K. Gupta and P. R.K, "Survey on Virtual Machine Placement Techniques in Cloud Computing Environment," Aug. 31, 2014. doi: 10.5121/ijccsa.2014.4401.
[21] U. Bellur,C. S. Rao,S. D. M. Kumar, "Optimal Placement Algorithms for Virtual Machines." Oct. 2023.
[22] Y. Song, X. Zhao, B. Wang, and Y. Sun, "Trading-Off Computing and Cooling Energies by VM Migration in Data Centers," Aug. 31, 2018, Institute of Electronics, Information and Communication Engineers. doi: 10.1587/transinf.2017edp7329.
[23] S. Lee et al., "Validating Heuristics for Virtual Machines Consolidation," Jan. 01, 2011.
[24] R. Dittner and D. Rule, "An Introduction to Virtualization," in Elsevier eBooks, Elsevier BV, 2007, p. 1. doi: 10.1016/b978-1-59749-217-1.00001-0.
[25] Fahimeh Farahnakian, University of Turku, Turku, Finland, fahimeh.farahnakian@utu.fi, Tapio Pahikkala, University of Turku, Turku, Finland, tapio.pahikkala@utu.fi, Pasi Liljeberg, University of Turku, Turku, Finland, pasi.liljeberg@utu.fi, Juha Plosila, University of Turku, Turku, Finland, juha.plosila@utu.fi, "Hierarchical Agent-Based Architecture for Resource Management in Cloud Data Centers." Jun. 2014.
[26] S. Garg, C. S. Yeo, A. Anandasivam, and R. Buyya, "Energy-Efficient Scheduling of HPC Applications in Cloud Computing Environments," Jan. 01, 2009, Cornell University. doi: 10.48550/arxiv.0909.1146.
[27] D. Laganá, C. Mastroianni, M. Meo, and D. Renga, "Reducing the Operational Cost of Cloud Data Centers through Renewable Energy," Sep. 27, 2018, Multidisciplinary Digital Publishing Institute. doi: 10.3390/a11100145.