Power Reduction Methods in a Cloud Computing Environment

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

18-11-2021 / 23-11-2021

CITATION

Segner, Brent (2021): Power Reduction Methods in a Cloud Computing Environment. TechRxiv. Preprint.
<https://doi.org/10.36227/techrxiv.17041673.v1>

DOI

[10.36227/techrxiv.17041673.v1](https://doi.org/10.36227/techrxiv.17041673.v1)

Power Reduction Methods in a Cloud Computing Environment

Brent Segner

Abstract

With the explosive growth of cloud computing over the last two decades, there has been a corresponding increase in data center power requirements to meet those needs (Geist & Reed, 2016). While there are numerous additional aspects of data center power management that could be explored from an environmental and facilities perspective, this research study will primarily focus on the cloud computing infrastructure. The objective of this focus will be on answering the research question “What methods are available to optimize power consumption in a cloud computing environment?”

Introduction

Over the last two decades, there has been a steady increase in the need for applications with a higher computer system performance and lower latency. This trend is expected to continue, if not accelerate, into the future with the introduction of 5G and Mobile Edge Computing (MEC) into the mobile computing ecosystem (Gu et al., 2020). These new technologies unlock many Internet of Things (IoT), Internet of Vehicles (IoV), tactile internet, artificial intelligence, augmented reality, and other intelligent and immersive experiences. While the products and services of this shift in mobile technology are virtually endless, all of these breakthroughs share similar underlying requirements for high-performance computing at extremely low latency (Shafi et al., 2017).

For either the technological or financial premise behind 5G / MEC to work, all services must be supported within a cloud computing environment located geographically close to the users (Palattella et al., 2016). While there has been a tremendous amount of progress in the virtualization space over the past few decades, the scale and performance tolerances required to support this evolution will surpass what was previously accomplished. Aside from simply solving the technical challenges of Software Defined Networking (SDN), Network Slicing, and Network Function Virtualization (NFV) (Shafi et al., 2017),

telecommunications operators will need to figure out how to operate a high-performance geographically dispersed cloud at scale. This challenge will include more than simply overcoming the technology constraints around latency and compute performance as these highly distributed locations have facility constraints with respect to the amount of space and power that they can accommodate.

Background: Power and Space Challenges for a Highly Distributed Cloud

As a prerequisite to addressing the challenges of power management within a highly distributed cloud, it is important to first have an understanding of both the broader shift towards cloud computing and the impact that it has had on power and cooling within the data center infrastructure. While the authors Geist and Reed (2016), Kwon (2020), and MirhoseiniNejad et al. (2020) all acknowledge this exponential rise in power consumption, each agrees that it is not sustainable in the long term for different reasons. In the case of Geist and Reed (2016), the authors state that with commercial data centers growing to 25-60 megawatts that the cost and facility design implications are reaching the limits of what can be sustained with current technology. While MirhoseiniNejad et al (2020) only lightly touch on datacenter design constraints by acknowledging

the power efficiency of the cooling system is a significant concern, the author surmises that the growing operations costs of supporting the power load should be considered a long-term limitation to supporting a distributed cloud at scale.

Although Kwon (2020) agrees with Geist and Reed (2016) and MirhoseiniNejad et al. (2020) about the unsustainable rise in data center power consumption, his rationale for the limiting factor is different. His focus is predominately on the environmental impacts of continued power growth at the current rates. Kwon (2020) estimates that data center power consumption in the United States has doubled since 2014 to 140 billion kilowatt-hours in 2020 which not only carries the operational costs that Geist & Reed (2016) were concerned with but also contributes 100 million metric tons of carbon into the atmosphere annually. These carbon emissions from data centers in the United States account for more than 4% of the total global carbon emissions during the same period.

Data Center Power Management Techniques

While supporting the power and cooling requirements of these data centers presents a formidable challenge for cloud operators to manage, there are numerous mechanisms that can

be employed to reduce and manage the concerns. This research discussed in the following pages will break down into three general categories of power management techniques:

- **Physical** - These techniques are intended to manipulate the settings of the physical servers (processor frequency, min/max power, CPU sleep state, etc.) to optimize power utilization.
- **Logical** - These techniques involve the placement and migration of virtualized workloads onto the hosts to consolidate the demand on as few servers as possible while increasing the overall environment utilization.
- **Application**- The utilization of application-specific profiles that facilitate the dynamic assignment of performance requirements when the virtualized workloads are provisioned into the cloud environment.

In Table 1, various techniques for data center power management are discussed. The benefits, shortcomings, and explanations for each technique are explained from related research studies.

Table 1

Current approaches to data center power management

Technique	Explanation	Benefits	Shortcomings	Studies
Virtualization	Dynamically provision resources	Breaks the traditional one application, one server model	If designed for peak loads, idle and underutilized resources can be expected.	(Singh et al., 2017) (Liu et al., 2014)
Server Consolidation	Reduce active servers by consolidating workloads	Increases server utilization. Vacated servers can be powered down.	Reduces server segregation and increases failure domain	(Liu et al., 2014) (Batista et al., 2015)
Dynamic Capacity Planning (DCP)	Automatically scale resources to meet demand by turning servers on and off	More energy efficient than dynamic power management (DPM)	Potential risk to application SLA if resources cannot be scaled quickly	(Singh et al., 2017) (Batista et al., 2015)
Load Balancing	Balance workload across different servers to increase utilization	Balances server utilization to reduce data center hotspots	Difficult to implement in a heterogeneous cloud environment	(Zakarya, 2018)(Goiri et al., 2012)
Scheduling & VM Placement	Place VM's onto a suitable (most energy efficient) server	Reduced consumption through the utilization of energy-efficient servers (DVFS)	Increased difficulty from a capacity planning and live migration perspective	(Yan et al., 2018) (Nzanywayin goma &

				Yang, 2017)
Live Migrations	Migrate VM's from over-utilized & underutilized to efficient servers	Less energy consumption	The service level of applications potentially impacted during migrations	(Yan et al., 2018) (Kumar & Sharma, 2018)
Renewables	Migrate VM's to servers using renewable power & cooling	Energy-efficient and economical	Renewables are intermittent and only available in specific geographies	(MirhoseiniN ejad et al., 2020) (Li et al., 2018) (Kwon, 2020)
Self Configuration or Adaptation	Energy management through the self-adaptive intralayer methodology	Dynamic energy management policies across all infrastructure layers	The capability to implement dynamic policies requires a deep understanding of infrastructure capabilities and application QoS requirements	(Yan et al., 2018) (Li et al., 2011)

Note: Table 1 describes the various approaches employed within a data center/cloud computing environment to reduce the overall power and cooling requirements. These approaches span the domains of physical, logical (i.e. virtualization layer), and application workload-based solutions.

Power and Space Management Research Contribution

This work contributes to a research focus on the application of power management techniques in smaller clouds that support high-performance workloads. The larger body of power management research has been oriented towards the power and cooling infrastructure within petascale data centers than techniques and methodologies that can be applied to the servers and workloads within the infrastructures. Research specifically in the area of power management techniques within smaller clouds will help facilitate cloud operators to manage the next generation of compute requirements in geographically distributed locations.

Theme 1: Power Optimization through Tuning of the Physical Infrastructure

Depending on several factors, including the size of the environment, technologies available, and type of workloads present, cloud operators may employ a combination of these to achieve the

greatest amount of power efficiency. Even within the largest cloud computing environments, average server utilization rates are less than 20% (Nzanywayingoma & Yang, 2017), which leaves a significant opportunity for power savings to be achieved through server tuning. The server power and performance tuning techniques generally fall under the category of Dynamic Power Management. Given the complexity around server power management, Kumar and Saxena (2017) and Zakarya (2018) explored four different categories that are broadly supported in modern servers and onboard processors.

- Dynamic Component Deactivation (DCD)
- Dynamic Performance Scaling (DPS / DFS)
- Dynamic Voltage Scaling (DVS)
- Dynamic Voltage and Frequency Scaling (DVFS)

Traditionally, many DPM components already exist within the base software of either the hardware or virtualization layer. While each of these approaches offers their own set of benefits, they are largely limited to specific use cases and come with their own restrictions. Table 2 below provides a

high-level view of each of the primary mechanisms for server power management and considerations for how they should be employed.

Table 2

Server Power management Techniques

	General Description	Use Case	Benefits	Restrictions
Dynamic Component Deactivation	Powering down of specific unused elements within a server	Unused processors can be powered down before server brought fully into service	Reduction in power requirements for unused hardware	Difficult to use once production workloads added into an environment
Dynamic Performance Scaling (DPS / DFS)	Establishing a min/max processor frequency for the server. Server/processors will regulate based on thresholds.	Create a slice of the environment with established performance parameters for specific workload.	Reduction in power & performance variability.	Infrastructure must be carved up in advance and specific workloads that match the performance specifications must be placed onto that hardware.
Dynamic Voltage Scaling (DVS)	Establishing a min/max wattage draw for the server. Server will regulate between performance thresholds.	Create a slice of the environment with predictable power draw requirements.	Reduction in power & performance variability.	Infrastructure must be carved up in advance and specific workloads that match the performance specifications must be placed onto that hardware.
Dynamic Voltage and Frequency Scaling (DVFS)	Server / processors will use a predefined algorithm to moderate performance and power draw.	Servers in cloud environments will regulate processor frequency and voltage requirements based on perceived workload.	Reduction of power during low utilization time periods.	Significant potential for high performance workloads to experience latency during sudden increase in demand.

Note: Table 2 represents the various approaches toward power management that may be employed upon a physical server. Each of these serves a specific purpose and set of use cases.

The most challenging power and performance management approach is the DVFS. DVFS is a mechanism in the latest generation of processors that allows for the control of CPU performance using a capability known as P-states. With the understanding that changing the CPU

performance through processor frequency adjustments can influence power draw, the core concept behind DVFS is identifying the optimum frequency to reduce power while not impacting the application workload that the server supports. As Table 2 illustrates, when evaluating DVFS, there

are a number of different mechanisms that should be considered when attempting to come to an understanding of the factors that influence performance. While the DVFS capability has a tremendous amount of potential, researchers Hsieh et al. (2020), Nzanywayingoma and Yang (2017), and Bini et al (2009) recognize the performance risks associated with its use. In trial testing, each of the authors acknowledges the potential for power savings at the risk of exceeding application latency tolerances.

Theme 2: Power Optimization through the Virtualization Layer.

While researchers and hardware manufacturers have focused on providing the ability to employ energy-saving techniques on the physical servers, there has been a parallel effort on

creating efficiencies in the virtualization layer. There are several different avenues of thought for how these efficiencies can be achieved within the virtualization layer. At a high level, the mechanisms either involve allocating the appropriate number and type of resources to satisfy the performance requirements or consolidating the workloads (Nzanywayingoma & Yang, 2017). The appropriate allocation of resources typically results in less capacity being consumed while the consolidation allows for one of the aforementioned techniques in the Server Power Management section to be employed (Geist & Reed, 2016). Table 3 lays out some of the high-level methods that the virtualization layer can be used for power optimization.

Table 3

Virtualized Resource Management Techniques

	Use Case	Benefit	Restrictions
Resource Allocation	Selecting the correct number and quantity of resources required to meet performance requirements.	Workloads are either oversized, wasting resources, or undersized creating performance issues. Right-sizing the resource allocation reduces these concerns.	The correct requirements must be identified to meet QoS SLA's in a particular environment.
VM Placement	New VM's are provisioned onto specific hosts or areas of the environment to satisfy performance, resilience, and infrastructure efficiency requirements.	Infrastructure capabilities are maximized and the environment can be tuned for the greatest amount of efficiency	Virtualized workloads are volatile and will frequently change which results in the initial provisioning potentially not remaining optimal after a period of time.
VM Migration	Once a VM is created and carrying a production load, it can be moved to a new physical host to optimize the utilization of the environment.	The workload migration to a new server or region within the cloud allows for optimization techniques referenced in the Server Power Management section to be employed.	The workload migration itself required additional power resources and there is often a performance disruption during the move process.

Note: Table 3 represents the different techniques that can be employed to manage virtualized workloads within a cloud environment.

Resource Allocation

Although it is less often thought of with respect to power reduction, virtual resource allocation plays a significant role in the process. In the research by Kecskemeti et al. (2012) and Li, Li, et al. (2018), the authors investigate a number of different operational impacts related to incorrect VM sizing. These studies agree with the research by Batista et al. (2015), that the majority of incorrectly sized VM's trend towards being oversized and therefore consume significantly more resources than required to maintain their QoS requirements. This overestimation of client demand not only leads to higher service costs but also underlying infrastructure and power costs. The dynamic performance evaluation model proposed by Batista et al. (2015) to identify the correct number of VM's, virtual cores (vCPU), disk size, and RAM needed to meet QoS requirements is actually implemented within the testing performed by Li, Li, et al. (2018).

VM Placement

Similar to the Resource Allocation, the initial VM placement has no shortage of research proposing different methodologies. While there are dozens of methods for VM placement, at the simplest level, they are all extensions of the classic bin packing problem. In the bin packing problem, the objective is to place as many variable sized objects into a fixed size bin as possible. In the context of virtualization, the aim is to place numerous VM's with variable amounts of CPU, RAM, and Network requirements onto a physical server with a fixed capacity.

This bin packing problem is a difficult problem to solve in a virtualized environment. One of the reasons is that the issue is multidimensional in that the VM's must be fitted onto a host that meets their CPU, RAM, and Network requirements within the CPU, RAM and Network limitations of the server. Since all resources on the server are shared resources, anything that has been allocated to another VM on the same server is no longer eligible for utilization (S & Nair, 2019). This problem is then scaled across the environment as "x" VM are fitted

onto "y" servers resulting in "z" potential solutions as represented by the formula $x * y = z$.

Considerations of Virtual Machine Placement

Aside from the physical fit of VM's onto a virtualized server, there are several additional considerations that must be accounted for. Although not an exhaustive list, these additional factors traditionally include:

- Performance (QoS) - Maintaining minimum performance QoS
- Cost - Total capital investment required to maintain VM on the server
- Latency - Proximity to users or adjacent workloads
- Reliability & availability - Support application and infrastructure resilience

Given the heavy utilization of public clouds over the last decade, the majority of studies as Table 4 illustrates have been focused on aspects of QoS and Cost. In the case of QoS, if the application does not perform within the acceptable tolerances, there is a direct impact on the downstream users and systems. Although it is not as meaningful to the end-users, inefficient architectures or allocation of resources drives additional costs into the business that can adversely influence the financial performance of the company. There are different philosophies among researchers on how to address this need to have insight into the actual QoS requirements while reducing costs. The researchers Sun et al. (2014) and Batista et al. (2015) advocate for the implementation of a uniform set of performance benchmarks across clouds and the implementation of industry-recognized standards like TOSCA to make better initial decisions. Taking a different approach, Hsieh et al. (2020) appear to concede that it will be difficult to correctly predict the workload performance in an environment ahead of implementation. He instead advocates for data-driven resource assessments and workload placement adjustments after implementation. The various considerations of virtual machine

placement within research studies are examined in Table 4.

Table 4

Virtual Machine Placement Studies

	QoS	Cost	Latency	Reliability
(Batista et al., 2015)	✓	✓	X	X
(Thinh et al., 2017)	✓	X	✓	X
(Nzanywayingoma & Yang, 2017)	✓	✓	X	X
(Basaran et al., 2010)	✓	X	X	✓
(Goiri et al., 2012)	✓	X	X	X
(Wang et al., 2017)	✓	X	X	X
(Yan et al., 2018)	X	X	X	✓
(Kumar & Sharma, 2018)	✓	X	X	✓
(Sun et al., 2014)	✓	✓	✓	✓
(Liu et al., 2014)	✓	✓	X	X
(Zakarya, 2018)	X	✓	X	X
(Hsieh et al., 2020)	✓	✓	X	X

Note: Table 4 illustrates the four primary considerations when evaluating the type and number of resources that a virtual machine will consume as well as logical and geographic placement. Each of these elements will have a direct impact on the overall performance and stability of the virtualized workload.

With the primary vehicle to reduce cost by increasing the utilization levels of current virtualized workloads, there is a considerable amount of effort expended towards understanding how heavily servers can be loaded with virtualized workloads before the QoS begins to degrade. In the articles by Batista et al. (2015), Nzanywayingoma and Yang (2017), Sun et al. (2014), and Hsieh et al. (2020) the researchers identify specific metrics and measurements that are indicative of detecting environmental stress.

Aside from studies similar to Thinh et al. (2017) and Sun et al. (2014), there has been little emphasis in current studies on the potential impacts of latency. These studies were mainly focused on the need for greater computational performance to offset geographical latency between applications. This focus is likely to shift in

the coming years with 5G and MEC as discussed by Nowak et al. (2021) and Ge et al. (2016).

Virtual Machine Migration

Once workloads are placed into an environment, the conditions in which they were initially placed have a tendency to change. This is the result of increased/reduced occupancy on the server, changes to the environmental configurations, and other performance-impacting considerations. In order to optimize the utilization of the environment or enhance QoS for the existing services once these changes occur, the utilization of a virtual machine migration can be a valuable tool. These migrations allow for the movement of

virtualized workloads from one physical host to another.

The primary considerations around virtual machine migration remain similar to those of the initial placement. Any migrations that are performed must account for the bin packing scenarios of the initial placement while at the same time accounting for any service disruption or performance degradation that would occur through the migration process. Aside from the placement and workload performance impacts, live migrations require a temporary increase in power consumption to facilitate resource placement.

Theme 3: Power Optimization using Application Performance Profiles.

One of the tremendous advantages of the cloud computing environment is providing the illusion of a virtually unlimited pool of resources to applications. This resource pool allows applications to scale up rapidly to meet demand and reduce resource requirements when they are no longer necessary. (Kumar & Sharma, 2018) The challenge for cloud providers is offering this same service to

many applications while ensuring no degradation in service or data loss. In order to accomplish this balancing act, the providers need to be able to guarantee a minimum Quality of Service (QoS) for the workloads running in the environment. This QoS refers to creating performance benchmarks that quantitatively measure how efficiently services in the infrastructure perform (Batista et al., 2015).

When applications look to deploy their services into a cloud, they have to select both the quality and quantity of resources that will be required (Li et al., 2011). Traditionally, the virtual machines responsible for executing the client's requests are arranged into classes based on their various memory configurations, virtual cores, and disk size. Researchers Goiri et al., (2012) and Shafiabadi et al., (2021) both believe that in addition to resource sizing, the underlying hardware profile that includes attributes such as CPU or GPU model should additionally factor into the classification. When applications select the appropriate class of virtualized infrastructure, theory must balance the need for performance against cost. Table 5 illustrates the most common mechanisms employed to determine whether a suitable number of resources have been assigned to support the current workload.

Table 5: CPU Performance Considerations

	Multi Threading	Frequency	Instruction Type	Compiler	Memory
(Thinh et al., 2017)	X	✓	X	X	X
(Matick et al., 2001)	✓	X	X	X	✓
(Dolbeau, 2017)	✓	✓	✓	✓	X
(Bini et al., 2009)	X	✓	X	X	X
(Wang et al., 2019)	X	✓	✓	✓	✓
(Liu et al., 2014)	X	✓	X	X	X

Note: Table 5 represents different elements that may impact the perceived CPU performance that a virtualized workload would experience.

If the resources selected cannot meet their QoS requirements, there is a significant potential for service impairment which can range from minor delays to a complete outage. On the other hand, if

the application selects the wrong combination of resources or overestimates its requirements, there is the possibility that it will incur much higher costs (Wang et al., 2017). This selection process is

difficult for applications that do not have a robust understanding of their specific requirements and

can be even more cumbersome if there is much variability in workload demand.

Table 6

Methods to detect host overloading

Type	Static utilization threshold based algorithms	Adaptive utilization based algorithms	Regression based algorithms
Explanation	Based on fixed CPU utilization threshold	Based on a statistical analysis of historical data of VM	Based on an estimation of future CPU utilization
Pros	Simple	Suitable for the dynamic environment (robust)	Better predictions of host overloading
Cons	Unsuitable for dynamic environment	Poor prediction of host overloading	Complex,
Examples	THR (Averaging threshold-based algorithm)	MAD (Median Absolute Deviation), IQR (Inter Quartile Range)	LR (Local Regression), LRR (Robust local Regression)

Note: With the emphasis on ensuring that application workloads can work within their QoS tolerances in a virtualized environment, there are a variety of measurements available. Table 6 illustrates five of the most common methods for evaluating whether a host can adequately support the workload.

To reduce cost without impacting performance, the application needs to have a well-grounded understanding of both its runtime characteristics and an understanding of the environment (Yan et al., 2018). In addition, depending on the deployment, it may be necessary to understand other applications running on the same infrastructure and to account for any potential interactions. This understanding between application performance requirements and infrastructure capability will provide the most efficient mechanism to maintain a minimum level of performance while reducing the total resource usage and cost.

In 2014, data centers within the United States consumed 70 billion kWh which amounted to just over 1.8 percent of the total domestic power consumption. As of 2020, the total data center power consumption had doubled to 140 billion kWh and almost 4 percent of the total domestic power consumption (MirhoseiniNejad et al., 2020). With public cloud providers experiencing an increase of electricity demand of approximately 20-fold, almost all of this growth in power consumption has been focused on expanding cloud computing infrastructure (Li et al., 2018). Until recently, since the growth had been primarily centralized in a few hyper-scale data centers, it was possible to manage the increase in power requirements by manipulating the data center infrastructure. With the current trend towards decentralization of cloud infrastructure, it becomes more imperative that cloud operators have the ability to answer the

Conclusion

Implications

research question regarding “What methods are available to optimize power consumption in a cloud computing environment?”

This trend of decentralization combined with cloud computing growth will continue to accelerate with the rise of 5G and MEC (Gu et al., 2020). These services will largely be located in smaller decentralized facilities that support specific local geographies. With these smaller locations, it will no longer be practical or economically feasible to take advantage of some of the techniques that have worked in the past to mitigate the power load on the cooling infrastructure. As Table 6 illustrates, the power efficiencies required to support these services will come through a deeper understanding of both the application and the infrastructure. There will need to be a holistic view within each

environment of how to meet the application's QoS needs while reducing the overall power and space footprint.

Gaps in Literature

Within each of the articles, various aspects of server power management, virtualized workload placement, or application performance are evaluated individually. There is little mention within the articles of the implications of applying more than one of these techniques to address the broader concerns of power management within a cloud environment. Given the diverse nature of the environments and the workloads that run within them, it is unlikely that the implementation of a single approach would fully address the power management needs of the cloud operator.

References

- Basaran, C., Suzer, M. H., Kang, K.-D., & Liu, X. (2010). Robust fuzzy CPU utilization control for dynamic workloads. *Journal of Systems and Software*, 83(7), 1192–1204. <https://doi.org/10.1016/j.jss.2010.01.031>
- Batista, B. G., Estrella, J. C., Ferreira, C. H. G., Filho, D. M. L., Nakamura, L. H. V., Reiff-Marganiec, S., Santana, M. J., & Santana, R. H. C. (2015). Performance Evaluation of Resource Management in Cloud Computing Environments. *PLoS ONE*, 10(11). <https://doi.org/10.1371/journal.pone.0141914>
- Bini, E., Buttazzo, G., & Lipari, G. (2009). Minimizing CPU energy in real-time systems with discrete speed management. *ACM Transactions on Embedded Computing Systems*, 8(4), 1–23. <https://doi.org/10.1145/1550987.1550994>
- Dolbeau, R. (2017). Theoretical peak FLOPS per instruction set: a tutorial. *The Journal of Supercomputing*, 74(3), 1341–1377. <https://doi.org/10.1007/s11227-017-2177-5>
- Ge, X., Tu, S., Mao, G., Wang, C.-X., & Han, T. (2016). 5G Ultra-Dense Cellular Networks. *IEEE Wireless Communications*, 23(1), 72–79. <https://doi.org/10.1109/mwc.2016.7422408>
- Geist, A., & Reed, D. A. (2016). A survey of high-performance computing scaling challenges. *The International Journal of High Performance Computing Applications*, 31(1), 104–113. <https://doi.org/10.1177/1094342015597083>
- Goiri, Í., Julià, F., Fitó, J. O., Macías, M., & Guitart, J. (2012). Supporting CPU-based guarantees in cloud SLAs via resource-level QoS metrics. *Future Generation Computer Systems*, 28(8), 1295–1302. <https://doi.org/10.1016/j.future.2011.11.004>
- Gu, X., Zhang, G., & Cao, Y. (2020). Cooperative mobile edge computing-cloud computing in Internet of vehicle: Architecture and energy-efficient workload allocation. *Transactions on Emerging Telecommunications Technologies*, e4095. <https://doi.org/10.1002/ett.4095>

- Hsieh, S.-Y., Liu, C.-S., Buyya, R., & Zomaya, A. Y. (2020). Utilization-prediction-aware virtual machine consolidation approach for energy-efficient cloud data centers. *Journal of Parallel and Distributed Computing*, 139(139), 99–109.
<https://doi.org/10.1016/j.jpdc.2019.12.014>
- Kecskemeti, G., Terstyanszky, G., & Kacsuk, P. (2012). Virtual Appliance Size Optimization with Active Fault Injection. *IEEE Transactions on Parallel and Distributed Systems*, 23(10), 1983–1995.
<https://doi.org/10.1109/tpds.2011.309>
- Kumar, K., & Saxena, V. (2017). Optimization of Energy Consumption for Task Scheduling on Uni-Processor and Multiprocessor for Step Topology under Distributed Environment. *International Journal of Computer Applications*, 161(8), 10–16.
<https://doi.org/10.5120/ijca2017913244>
- Kumar, M., & Sharma, S. C. (2018). Deadline constrained based dynamic load balancing algorithm with elasticity in cloud environment. *Computers & Electrical Engineering*, 69(69), 395–411.
<https://doi.org/10.1016/j.compeleceng.2017.11.018>
- Kwon, S. (2020). Ensuring renewable energy utilization with quality of service guarantee for energy-efficient data center operations. *Applied Energy*, 276(276), 115424.
<https://doi.org/10.1016/j.apenergy.2020.115424>
- Li, A., Zong, X., Kandula, S., Yang, X., & Zhang, M. (2011). CloudProphet. *ACM SIGCOMM Computer Communication Review*, 41(4), 426.
<https://doi.org/10.1145/2043164.2018502>
- Li, H., Li, W., Feng, Q., Zhang, S., Wang, H., & Wang, J. (2018). Leveraging content similarity among VMI files to allocate virtual machines in cloud. *Future Generation Computer Systems*, 79(79), 528–542.
<https://doi.org/10.1016/j.future.2017.09.058>
- Li, X., Garraghan, P., Jiang, X., Wu, Z., & Xu, J. (2018). Holistic Virtual Machine Scheduling in Cloud Datacenters towards Minimizing Total Energy. *IEEE Transactions on Parallel and Distributed Systems*, 29(6), 1317–1331.
<https://doi.org/10.1109/tpds.2017.2688445>
- Liu, Z., Xiang, Y., & Qu, X. (2014). Workload-Aware and CPU Frequency Scaling for Optimal Energy Consumption in VM Allocation. *Mathematical Problems in Engineering*, 2014, 1–12.
<https://doi.org/10.1155/2014/906098>
- Matick, R. E., Heller, T. J., & Ignatowski, M. (2001). Analytical analysis of finite cache penalty and cycles per instruction of a multiprocessor memory hierarchy using miss rates and queuing theory. *IBM Journal of Research and Development*, 45(6), 819–842.
<https://doi.org/10.1147/rd.456.0819>
- MirhoseiniNejad, S., Moazamigoodarzi, H., Badawy, G., & Down, D. G. (2020). Joint data center cooling and workload management: A thermal-aware approach. *Future Generation Computer Systems*, 104(104), 174–186.
<https://doi.org/10.1016/j.future.2019.10.040>
- Nowak, T. W., Sepczuk, M., Kotulski, Z., Niewolski, W., Artych, R., Bocianiak, K., Osko, T., & Wary, J.-P. (2021). Verticals in 5G MEC-Use Cases and Security Challenges. *IEEE Access*, 9, 87251–87298.
<https://doi.org/10.1109/access.2021.3088374>
- Nzanywayingoma, F., & Yang, Y. (2017). Efficient Resource Management techniques in Cloud Computing Environment: A Review and discussion. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 15(4), 1917.
<https://doi.org/10.12928/telkomnika.v15i4.6574>
- Palattella, M. R., Dohler, M., Grieco, A., Rizzo, G., Torsner, J., Engel, T., & Ladid, L. (2016). Internet of Things in the 5G Era: Enablers, Architecture, and Business Models. *IEEE Journal on Selected Areas in Communications*, 34(3), 510–527.
<https://doi.org/10.1109/jsac.2016.2525418>
- S, K., & Nair, M. K. (2019). Bin packing algorithms for virtual machine placement in cloud computing: a review. *International Journal of Electrical and Computer Engineering (IJECE)*, 9(1), 512.
<https://doi.org/10.11591/ijece.v9i1.pp512-524>

- Shafi, M., Molisch, A. F., Smith, P. J., Haustein, T., Zhu, P., De Silva, P., Tufvesson, F., Benjebbour, A., & Wunder, G. (2017). 5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice. *IEEE Journal on Selected Areas in Communications*, 35(6), 1201–1221. <https://doi.org/10.1109/jsac.2017.2692307>
- Shafiabadi, M., Pedram, H., Reshadi, M., & Reza, A. (2021). An accurate model to predict the performance of graphical processors using data mining and regression theory. *Computers & Electrical Engineering*, 90, 106965. <https://doi.org/10.1016/j.compeleceng.2020.106965>
- Singh, A. K., Dziurzanski, P., Mendis, H. R., & Indrusiak, L. S. (2017). A Survey and Comparative Study of Hard and Soft Real-Time Dynamic Resource Allocation Strategies for Multi-/Many-Core Systems. *ACM Computing Surveys*, 50(2), 1–40. <https://doi.org/10.1145/3057267>
- Sun, L., Dong, H., Hussain, F. K., Hussain, O. K., & Chang, E. (2014). Cloud service selection: State-of-the-art and future research directions. *Journal of Network and Computer Applications*, 45(45), 134–150. <https://doi.org/10.1016/j.jnca.2014.07.019>
- Thinh, T. Q., Tang, J., La, Q. D., & Quek, T. Q. S. (2017). Offloading in Mobile Edge Computing: Task Allocation and Computational Frequency Scaling. *IEEE Transactions on Communications*, 65(8), 1–1. <https://doi.org/10.1109/tcomm.2017.2699660>
- Wang, L., Qian, D., Luan, Z., Wei, G., Wang, R., & Yang, H. (2017). Speeding up profiling program's runtime characteristics for workload consolidation. *PLOS ONE*, 12(4), e0175861. <https://doi.org/10.1371/journal.pone.0175861>
- Wang, Y., Lee, V., Wei, G.-Y., & Brooks, D. (2019). Predicting New Workload or CPU Performance by Analyzing Public Datasets. *ACM Transactions on Architecture and Code Optimization*, 15(4), 1–21. <https://doi.org/10.1145/3284127>
- Yan, Y., Guo, P., Cheng, B., & Zheng, Z. (2018). An experimental case study on the relationship between workload and resource consumption in a commercial web server. *Journal of Computational Science*, 25(25), 183–192. <https://doi.org/10.1016/j.jocs.2017.05.019>
- Zakarya, M. (2018). Energy, performance and cost efficient datacenters: A survey. *Renewable and Sustainable Energy Reviews*, 94(94), 363–385. <https://doi.org/10.1016/j.rser.2018.06.005>